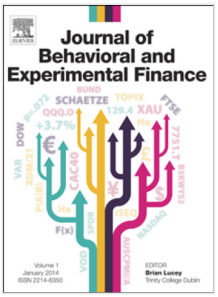
Accepted Manuscript



Social media big data and capital markets—An overview

Jaroslav Bukovina

PII: S2214-6350(16)30027-2
DOI: http://dx.doi.org/10.1016/j.jbef.2016.06.002
Reference: JBEF 83

To appear in: *Journal of Behavioral and Experimental Finance*

Received date: 12 May 2016
Revised date: 13 June 2016
Accepted date: 22 June 2016

Please cite this article as: Bukovina, J., Social media big data and capital markets—An overview. *Journal of Behavioral and Experimental Finance* (2016), http://dx.doi.org/10.1016/j.jbef.2016.06.002

Social Media Big Data and Capital Markets – an Overview

Jaroslav Bukovina

Department of Finance, Faculty of Business and Economics, Mendel university in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: jaroslav.bukovina@mendelu.cz.

Abstract

A growing body of research and practical applications employ social media data as the proxy for a complex behavior of a society. This paper provides an overview of academic research related to a link between social media and capital markets. The theoretical rationale of this relationship is predominantly defined by behavioral finance. Behavioral finance augments the standard model of efficient markets and considers less rational factors like investors' sentiment or public mood as influential for asset pricing and capital market volatility. In this context, social media is a novel tool enabling the collection of data about such less rational factors at the level of a society. The paper introduces social media data from a technical and economic point of view. In addition, it contributes to the theoretical construction of the transmission mechanism between social media and capital markets currently missing in the literature. Subsequently, the paper summarizes the main findings in this field and outlines future challenges in this research.

Key words

Social media, retail investors, information demand, sentiment, transmission mechanism

JEL: G02, G19

Acknowledgements

This work was supported by the Internal Grant Agency of PEF Mendelu (PEF_DP_2016010). The author would like to thank to anonymous reviewers and participants of Academy of Behavioral Finance Conference in Venice 2016 for valuable comments.

Introduction

Social media can be classified without dispute among the current phenomena in our society. Social media connects people from all over the world into one virtual community. The main goal of social media is the ability to facilitate communication and content sharing. However, for the purpose of this paper, there is another fact no less relevant: social media is a great database of society's behavior. Data provided by social media, so-called "big data", is very popular and many practical applications as well as academic research have been accomplished in this field. The goal is clear - to better understand the behavior of a society.

The purpose of this paper is to provide an overview of the current state of the art in research related to employment of social media big data in the field of capital markets. The main focus is to provide for inclusion in this growing research, summarize main findings and outline the challenges for further research. The contribution of the paper is a theoretical proposal of the transmission mechanism from social networks to capital markets. Despite the growing research, there has been no clear definition of this relationship yet. The rest of the paper is structured as follows. The next section provides the reader with an understanding of social media, its forms and functions. Further, it describes big data and sentiment analysis. Though these two sections go into deep technical detail, they are essential for understanding the challenges involved in data collection via social media. The third section presents the link between social media and capital markets from an economic point of view in the form of the transmission mechanism. The fourth section summarizes the contributions and findings of several important papers. The fifth section proposes the potential prospects of further research and the last section concludes.

1 Understanding social media

From a technical point of view, social media are web-based or mobile technologies necessary for operating of highly interactive platforms where users create, modify and share user-generated content (Kietzmann et al., 2011). Boyd and Ellison (2008) use instead of the term "social media" the term "social network site" and characterize it also as a web-based services which allow for users to create a public or semi-public profile and create a tree of connections with other users. Kaplan and Haenlein (2010) describe social media as internet-based applications that incorporate the ideas and technology of a previous web structure like Web 2.0 but allow for individual users to create and exchange web content.

Social media has evolved into many forms. Kaplan and Haenlein (2010) provide the classification scheme to sort social media in a systematic manner. However, for the purpose of this paper, this deeper classification is not relevant because the research related to capital markets uses predominantly the data from social media Facebook and Twitter. The choice of these two social media is most likely related to practical reasons like data availability, number of users or high popularity in the western world, because current research in this field examines predominantly the US or advanced European capital markets. Deeper understanding of the social media functioning is provided in the paper of Kietzmann et al. (2011), who constructed a framework of seven social media building blocks. This framework describes the social media environment and social media audience in detail. For the purpose of this paper, the building blocks "Sharing" and "Conversations" are the most important because they are the source of "big data". In other words, social media users lead conversations in many forms like Facebook "comments" or Twitter "tweets", and they are able to share them with others. Such data can be collected and further analyzed. However, these two blocks are results of other blocks, in particular "Presence", "Identity", "Groups", "Relationships" and "Reputation". Their functionalities and explanation are shown in Kietzman et al. (2011).

The big data provided by social media is not the only source of data employed in the analysis related to capital markets. Several papers presented below employ data from the Google search engine, so-called Google queries. To be specific, Google provides data about the search volume for a given phrase. Technically, the volume of a search phrase is publicly available (http://www.google.com/trends/) as an index – Search Volume Index (SVI). According to the index principals, data is available in values within the interval (0,100). A hundred value is the maximum reached volume of a specific search phrase for the given period in a comparison to all searches that have been done via Google at the given point in time. The rest of a time series is scaled according to this maximum. Zero represents an insufficient volume of searches. The volume of searches for a given phrase is available as of January 2004, usually on a weekly basis and exceptionally, in case of less popular searches, on a monthly basis. Moreover, search data can be filtered according to several criteria, e.g. geography, time or standard searches vs. searches for news. Google search queries are considered in this paper also as big data because it provides insight about the interest of a society in the searched topic. From a technical point of view, the Google search engine is not standard social media like Twitter and Facebook, but from a practical point of view, this discrepancy is not relevant for the purpose of this paper. In general, Google provides data about society's behavior as do both Twitter and Facebook.

1.1 Big data

The Web was originally structured as a "read-only" tool. Today it has transformed to "read and write" one (Cambria et al., 2013). In line with this idea, the previous part introduced social media as a platform, which enables users to create and share web content. However, in the context of the paper, social media must be understand as a medium - a source of big data only and big data (users' inputs) is the component that practical applications and academic research is mostly interested in.

In general, big data can be characterized as follows. Halevi and Moed (2012) define big data as the large sets of data, which cannot be processed by traditional management tools due to their size and complexity. Big data is high-volume, high-velocity and high-variety information that requires a cost-effective and innovative approach to processing data, which release the opportunity for enhanced insight, decision-making and process automation (Beyer and Laney, 2012; Gartner, 2015). The previous definitions need to be augmented in line with the idea of this paper. In particular, big data is large data sets related to the behavior of individuals and society. It has to be processed computationally due to its complexity. In the field of capital markets, social media data is applied to reveal the deeper insights, trends and associations between society and capital markets.

The field of finance is not the first to employ social media big data. It has been previously widely used to study various notable social events like predicting political elections (Yu et al., 2008; Tumasjan et al., 2010), to detect natural disasters (Sakaki et al., 2010) or disease epidemics (Ginsberg et al., 2008; Lamb et al., 2013) in a real-time. In commercial sphere, social media big data serves as an early "economic indicator" or an electronic tool for word-of-mouth marketing. In particular, Gruhl et al. (2005) and Mishne and Rijke (2006) show the affirmative evidence that social media activity can predict book and movie sales respectively. Choi and Varian (2012) forecast the automobile sales, unemployment claims, travel destination planning and consumer confidence based on Google queries. Jansen et al. (2009) discuss the implications of social media in corporate marketing strategy. Even the policy makers and market regulators are becoming interesting in the "power" of big data to reflect the behavior of a society. McLaren and Shanbhogue (2011) from Bank of England, D'Amuri and Marcucci (2012) from Bank of Italy or Saxa (2014) from Czech National Bank discuss the employment of Google queries as a real-time indicator for the labor and housing markets.

The previous paragraph presents the specific employment of social media big data in various fields of research and section 3 does as well in the field of finance. However, to properly evaluate existing empirical results and compare them across the papers, one has to be introduced to the sentiment analysis methods.

1.1.1 Sentiment analysis

Each definition of big data above incorporates the issue of difficulty in big data processing. Big data processing, in another words sentiment analysis, can be considered as one of the most important points in big data research. Sentiment analysis or opinion mining describes various computational techniques focused to discover, extract and distil the human emotions, feelings or opinions from textual information within the web content towards the certain entities (Fang and Zhan, 2015; Godsay, 2015). Today, sentiment analysis has evolved into an autonomous and currently popular field of study. In general, there are two methods for sentiment analysis. 1. The machine-learning method can be considered as a program that learns from data. This program works on the predefined model (classifier) that is fed with examples of text (tagged text, vector of feature values – e.g. positive or negative sentiment) that is later used to classify an untagged corpus of text. The initial definition of feature values is crucial to the final success of the machine-learning approach. An advantage of the machine-learning method lies in its ability to be defined in detail for specific contexts. On the contrary, a disadvantage is its low applicability to new data due to costly text tagging or unavailability of sufficient examples from which the machine-learning program can learn (Annett and Kondrak, 2008; Gonçalves et al., 2013; Godsay, 2015). 2. The lexical method utilizes a predefined (pre-tagged) list of words, lexicon or dictionary where each word or phrase conveys the specific sentiment. Studied corpus of text is then compared against the list of words usually by the algorithm. An advantage of this approach is that no pre-tagging is necessary, but it is a shortcoming when specific context needs to be analyzed because a unique dictionary can be unavailable (e.g. if slang is common in the studied corpus of text) (Annett and Kondrak, 2008; Gonçalves et al., 2013). The process of sentiment analysis consists of several steps. 1. Social media big data has to be downloaded. 2. Web content transformed to text content only. 3. Data transformation from a qualitative into a quantitative nature based on the machine-learning or lexical method. Further technical details of sentiment analysis process are provided in the research of Pang and Lee (2008), Russel (2011), Liu (2012) or Godsay (2015).

In the context of the paper, when one employs the data from Twitter (tweets) or Facebook (comments), the transformation from the qualitative nature (emotions, opinions) into the quantitative one is necessary based on the methods described above. However, one can avoid the performance of complex sentiment analysis via the already transformed social media big data. In particular, Google Corporation provides the data of Google queries introduced above and Facebook created the Gross National Happiness Index (FGNHI)[1]. The employment of Google queries is popular in various fields presented in the previous section due to public availability. However, this information is quantitative

[1] Unfortunatelly, FGNHI is not available anymore due to privacy concerns.

in nature only and qualitative distinction, whether people search for positive or negative information, is missing. Qualitative distinction was available by FGNHI, which provided the daily sentiment data for twenty countries. This index consisted of a number of anonymous positive and negative words employed in comments of people.

Previous paragraphs briefly introduced the technicalities behind the big data and sentiment analysis. In further sections, the author shows its rationale in the field of finance.

2 Economic rationale in social media big data applications

Social media big data captures the activity of individuals, interactions among them, or more precisely, the complex behavior of a society. Society's behavior and its relation to capital markets are a dominant part of analysis in the field of behavioral finance. Therefore, especially the behavioral finance framework serves as a main motivation for the employment of social media big data in the field of capital markets. In particular, behavioral finance challenges the notions of efficient markets and proposes the factors like animal spirits (Shiller, 1984), social mood (Nofsinger, 2005), investor sentiment (Baker and Wurgler, 2007) or psychological factors (Fenzl and Pelzmann, 2012) as a source of market volatility and anomalies. The previous research employed questionnaires (Case and Shiller, 2003) or derived sentiment proxy (Baker and Wurgler, 2007) to capture such factors. However, only social media bring the opportunity to collect detailed data about these factors at the aggregate level of a society.

The next important contribution of behavioral finance research is the realistic assumption about the existence of investors with bounded rationality (De Long et al., 1990, Shleifer and Vishny, 1997, Barberis et al., 1998). In reality, these less rational investors are especially the retail investors. In the future, one can anticipate an increasing number of retail investors due to technological development and growing number of trading platforms. These economic agents are a major part of the transmission mechanism from social media to capital markets. Their existence has two following economic interpretations in the current research.

2.1 Information demand

The first idea is based on the information demand of investors, so-called investors' attention. In particular, retail investors use investment forums of social media or the Google search engine as a publicly available source of information about capital markets because they have limited sources and access to professional databases like Bloomberg or Thomson Reuters. This idea is presented in papers of Da et al. (2011), Vlastakis and Markellos (2012), Ding and Hou (2015) who employ the Google

queries in their analysis. Information demand is also presented in Sprenger et al. (2014a, 2014b), who proposes the investment forums of Twitter as an alternative information source for retail investors. Investment forums on social media connect retail investors with a discussion about the securities, market and their fundamentals. The rationale of information demand is presented in Figure 1, which shows a simple transmission mechanism.

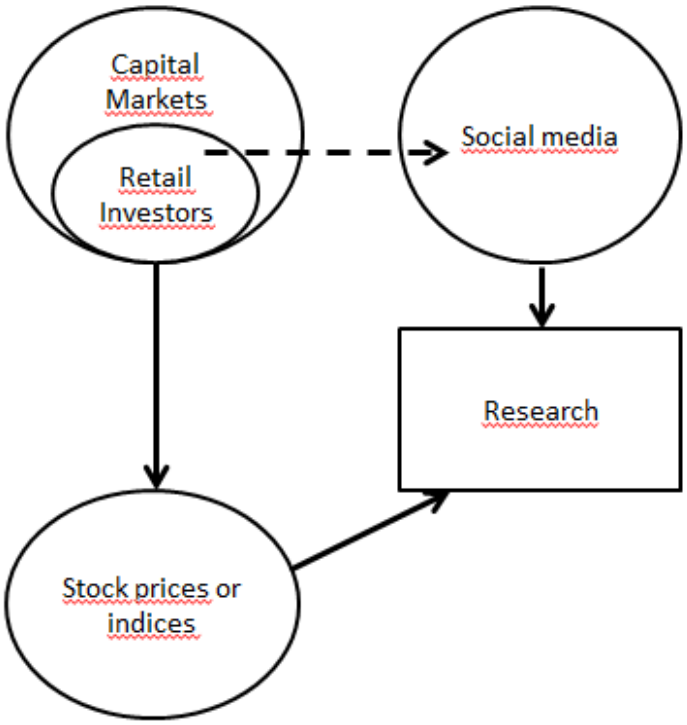


Figure 1. Transmission mechanism – Information demand; Source: (Author's work)

The transmission mechanism begins with capital markets and retail investors are a part of it. Retail investors, due to limited resources, use social media investment forums or the Google search engine as the main information source. However, we can assume only a partial reflection of information demand in social media data because social media are not the only source of information for retail investors. This is indicated in Figure 1 by the dashed arrow from retail investors to social media. Furthermore, capital market aggregate activity, which also includes trading of retail investors "defines" the value of stock prices or stock indices. This is shown by the solid arrow. Subsequently, research in this area employs the data from capital markets and social media to better understand the behavior of retail investors for stock price/market formation and its volatility. Important parts of the transmission mechanism are in circles. "Research" is in the rectangle, to distinguish it from the main parts of the transmission mechanism.

2.2 Sentiment of society

The second point of view is based on a reaction of society to existing information. To be specific, social media enable us to create, share and respond to existing information. Such a combination of reactions is a valuable source of data mostly about opinions, emotions or social mood shared by the social media audience. Such an audience consists especially of ordinary people who share their opinions, mood or emotions about concrete information. A good example is a corporate website on Facebook where ordinary people share their opinions related to the individual company. In financial literature, such information like opinions, moods and emotions can be described by the term sentiment, i.e. a set of irrelevant information not related to company fundamentals. Rich research (Kumar and Lee, 2006; Baker and Wurgler, 2007; Barber and Odean, 2011) implies that less-rational groups of investors are prone to behave according to sentiment. This idea is shown in the papers of Bollen et al. (2011), Mao et al. (2011) who study the Twitter data or Karabulut (2013), Siganos et al. (2014) and Bukovina (2015), who employs the data of Facebook. Figure 2 shows this transmission mechanism.

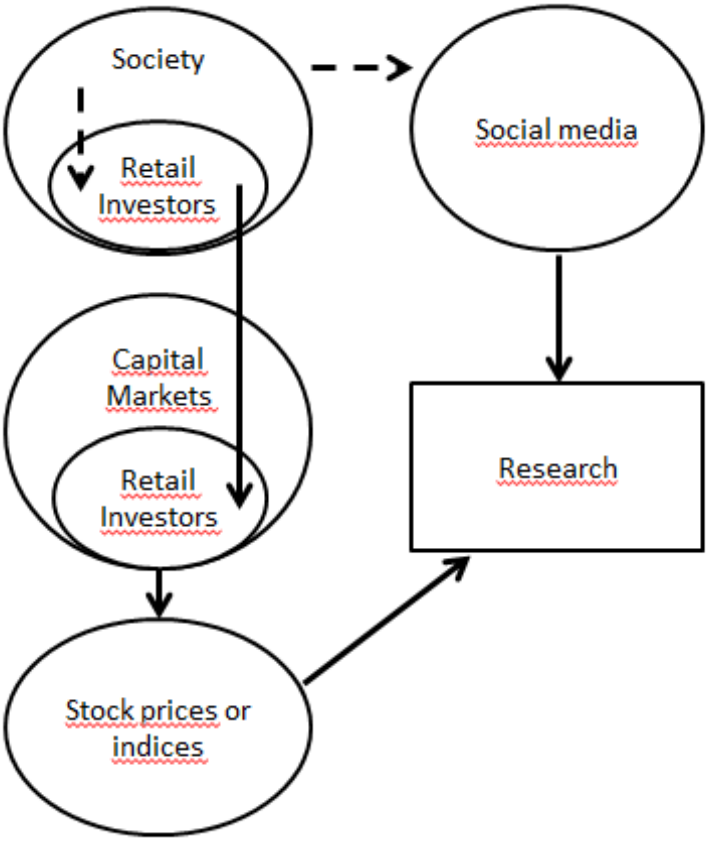


Figure 2. Transmission mechanism – Reaction of a society to information; Source: (Author's work)

At its beginning, there is the society and retail investors are a part of it. In the society, sentiment forms randomly due to many factors, which cannot be predicted. Retail investors as a part of a society can

be at least partially influenced by sentiment. They are aware of the fundamentals (signal), but the sentiment (noise) can be also the deciding factor. In Figure 2, it is presented by a dashed arrow because we cannot assume the influence of sentiment alone. Similarly, we can assume only a partial reflection of society's sentiment in social media shown by dashed arrow as well. The next part of transmission mechanism is the capital markets, and since retail investors are a part of it, there can be the impact of sentiment as well. Subsequently, the total trading of institutional and retail investors is reflected in the stock price, and if retail investors are influenced by sentiment, the stock price can also reflect this.

The most important contribution of this chapter is the understanding that social media is only a tool for how to technically capture the data about the behavior of a society. However, what really matters is the behavior of a society, the existence of retail investors and their influence on capital markets. Moreover, to employ the correct transmission mechanism, when one describes the link between society and capital markets, it is necessary to understand the rationale of social media big data described in above. Information demand would be the best represented by Google queries when researchers study the interest of retail investors via the search for stock tickers without the impact of sentiment. On the contrary, Twitter tweets of Facebook comments would be a better choice for researchers to evaluate the qualitative value of information and its impact to stock prices or capital markets.

3 Overview of the main empirical findings

The aim of this section is to document the main empirical findings about the link between behavior of society tracked by social media and by capital markets. To the best of the author's knowledge, the following subsections present the majority of existing papers, which examine the link between social media and capital markets based on theoretical argumentation and the transmission mechanism proposed above. However, there are many papers, which employ the social media data but do not focus on a better understanding of the capital market behavior. An example of such research is technically focused papers, which try to propose the new or advanced data-mining methods, and they test it on a link between social media and capital markets. Therefore, below subsections present only the papers, which follow the economic rationale and transmission mechanisms proposed above. All presented papers "agree" about the existence of less rational investors who bear the information demand or sentiment with subsequent reflection in their trading. The only disagreement is that if information demand reflects either less sophisticated investors and noise information (Da et al. 2011,

Joseph et al. 2012, Ding and Hou, 2015) or more informed investors and more efficient markets (Vozlyublennaia, 2014).

This section has three subsections based on the name of social media data, which is employed as the main source of big data. Each subsection is summarized in a table, which highlights the contribution of key social media papers mentioned in the given subsection. Tables I. – III. consist of the paper reference, employed data sample (either unique, processed by sentiment analysis or standardized and publicly available), transmission mechanism describing the economic rationale, the focus of the paper to the "macro" level of market or the "micro" level of securities and brief summary of major findings.

3.1 Facebook

The paper of Karabulut (2013) employs Facebook's Gross National Happiness Index (FGNHI) as a direct indicator of investor sentiment, and its result shows the ability of FGNHI to predict changes in the aggregate US market daily returns and trading volume. This result is significant even when controlled by past stock market volatility, daily economic conditions or the turn-of-the-year effect. Karabulut (2013) argues these results are consistent with the noise trader models because the positive impact of sentiment is fully reversed during the following two trading weeks. Siganos et al. (2014) provide the international evidence within the 20 markets between the daily sentiment and trading based on FGNHI data as well. Their main findings show positive contemporaneous relation to stock market returns. In addition, they suggest causality from sentiment to stock market because sentiment on Sunday affects the Monday stock returns. Siganos et al. (2014) explain their results based on a theory of behavioral finance proposed by De Long et al. (1990) that sentiment has an impact when retail investors are plentiful and there is the existence of limits in arbitrage. Karabulut (2013) and Siganos et al. (2014) focus on the effect of investor sentiment in the aggregate market. Bukovina (2015) is focused on a micro level because he studies the impact of sentiment on the sample of the US blue-chip companies. He employs the data from Facebook pages of individual corporations. Bukovina (2015) employs volumes of data without qualitative distinction to positive or negative. However, he considers such data as a sentiment because corporate profiles on Facebook "summarize" the behavior of ordinary people who discuss their opinions, emotions or expectations towards the corporation and its product or services. Such information cannot be considered as relevant fundamentals. However, based on the main findings, Bukovina (2015) is able to evaluate the qualitative nature of sentiment according to the results of panel regression analysis and a sign of regression coefficients. In particular, he suggests the existence of occasional peaks in Facebook sentiment with negative impact on stock returns. Bukovina (2015) offers the explanation about the bolder negative sentiment in comparison to a positive one based on the loss aversion theory (Kahneman and Tversky, 1979).

Table I. Key highlights - Facebook

Reference	Data sample	Transmission mechanism	Market/ Securities	Main Findings
Karabulut (2013)	Standardized (FGNHI)	Reaction of a society	Market	Sentiment predicts changes in daily returns and trading volume on the U.S. stock market; This link is positive and temporary; Reversal follows during the next weeks.
Siganos et al. (2014)	Standardized (FGNHI)	Reaction of a society	Market	Positive and contemporaneous link between daily sentiment and stock returns within 20 international markets; Causality – Sunday's sentiment affects Monday's stock returns; Negative sentiment increases trading volume and return volatility.
Bukovina (2015)	Unique (Company pages on Facebook)	Reaction of a society	Securities	The presence of occasional peaks in sentiment followed by negative impact on stock returns.

3.2 Twitter

One of the most famous papers in this field is Bollen et al. (2011) who derive the social mood from Twitter feeds and find the correlation as well as causality in Granger's sense in the stock market. The public sentiment has been extracted from approximately 10 million tweets posted by 2.7 million users. The Dow Jones Industrial Average (DJIA) represents the market in their study. Moreover, Bolen et al. (2011) define the 6 dimensions of social mood and the dimension "Calm" significantly improves the prediction of the DJIA. Mao et al. (2011) employ several sentiment sources to model the public and investors sentiment. Twitter feed is one of them. In particular, they constructed two mood indicators. 1. Twitter Investor Sentiment (TIS) is a ratio of bullish and bearish tweets. 2. Tweet Volumes of Financial Search Terms (TV-FST) is a tweet volume for search queries of various financial terms. Both these indicators are significant predictors for daily market returns. Both indicators outperform traditional sentiment tools like sentiment surveys (Investor Intelligence or Daily Sentiment Index) and news media sentiment. Moreover, TV-FST indicator is even more efficient in comparison to the indicator based on Google queries in their study. Sprenger et al. (2014a, 2014b) study the Twitter microblogging forums related to stock markets. Sprenger et al. (2014a) argue these forums contain news about stock prices and market only, without the noise, which is likely included in the general Twitter feed as is the case of Bollen et al. (2011). Sprenger et al. (2014a) are focused on the level of individual stocks because they analyze roughly 250,000 stock-related tweets on a daily basis. Their findings show the associations between tweet sentiment, stock returns and trading volume. Moreover, they show that the microblogging community can recognize the "extraordinary" users who consistently share high-quality investment advice. Similarly, Sprenger et al. (2014b) study company

events via Twitter microblogging forums. They identify good and bad news in a sample of more than 400,000 stock-related tweets. Therefore, they are able to capture the market reaction to news as well as to the sentiment due to the qualitative distinction of news. Their findings show the different behavior of the capital market according to categories of company events. In particular, categories like "Mergers and Acquisitions" or "Earnings" used to be a surprise for market participants in comparison to categories like "Joint Venture" or "Development" that rarely move the prices. Moreover, their results show that positive news is often leaked and incorporated into stock prices before the official announcement. On the contrary, negative news is predominantly surprising, related to the occurrence of an event within the day of occurrence.

Table II. Key highlights – Twitter

Reference	Data sample	Transmission mechanism	Market/ Securities	Main Findings
Bollen et al. (2011)	Unique (tweets)	Reaction of a society	Market	Mood dimension "Calm" significantly improves the prediction of DJIA index
Mao et al. (2011)	Unique (tweets)	Reaction of a society	Market	Twitter sentiment indicators predict daily market returns, outperform standard sentiment tools and even Google queries
Sprenger et al. (2014a)	Unique (tweets)	Information demand	Securities	Contemporaneous positive link between tweets bullishness and returns, messages volumes and trading volume;
Sprenger et al. (2014b)	Unique (tweets)	Information demand	Securities	Social media stock forums reliably identify stock-relevant news; Stock returns are more pronounced prior to good news than for bad news; Leaks of positive news before official announcements; Negative news as a surprise;

3.3 Google

One of the first papers examining Google data is Da et al. (2011), who propose the Google search queries of stock tickers as a new direct measure of investor attention. They accomplished the analysis in a sample of Russell 3000 Index stocks. According to their findings, the increased volume of searches for the studied sample of stocks predicts higher stock prices during the next 2 weeks, and eventual price reversal follows within the year. In addition, it predicts a company's IPO high first-day return and long-term underperformance. Da et al. (2011) explain their results according to the argument of Barber and Odean (2008) that retail investors' heightened attention increases the buying pressure on prices. Mao et al. (2011) mentioned in the previous section employ Google queries as well. Their findings show that Google queries are causal in the Granger sense with market index closing values and its volumes. Moreover, forecasting accuracy was improved by adding Google queries volumes,

especially when the market index moved downwards and market volatility is higher. Joseph et al. (2012) consider the online ticker searches as a proxy for investor sentiment. In addition, they suggest that a ticker search represents more likely a "buying" decision because an investor who considers "selling" already has the relevant information since he owns the specific security. However, they do not discuss the issue of short-selling where one does not own the specific security. Joseph et al. (2012) accomplished the analysis in a sample of S&P 500 companies formed in portfolios based on search intensity, returns from volatility and dual-sorted portfolios (combination of search intensity and volatility). Their findings are very similar to Da et al. (2011) because the previous search intensity forecasts abnormal returns and higher trading volumes. Moreover, their results are in line with Baker and Wurgler (2007) because sensitivity of returns to search intensity is positively related to the difficulty of a stock being arbitraged. Da et al. (2015) employed Google queries as an sentiment indicator. It is in "contrast" with a previous paper of Da et al. (2011) where Google queries served as a proxy for information demand. In particular, Da et al. (2015) constructed the Financial and Economic Attitudes Revealed by Search (FEARS) index. This index aggregates search queries from U.S. households like "recession", "bankruptcy" or "unemployment". It was employed to a model link with asset prices, volatility and fund flows at the level of portfolios and market level represented by the S&P 500 index. Their findings show that FEARS predict return reversals, temporary market volatility and fund flow from equity funds to bond funds. Da et al. (2015) interpret their results as consistent with the noise trading hypothesis of De Long et al. (1990). Ding and Hou (2015) conduct a study in the sample of the S&P 500 as Joseph et al. (2012) but with focus on the stock liquidity. Their results show that increased investor attention leads to a reduced relative bid-ask spread and the higher turnover rate. An explanation of their results is based on the "investor recognition hypothesis" (Grullon et al. 2004; Fang and Peress, 2009), which states that stocks with higher attention are more "recognized" and subsequently more liquid. Latoeiro et al. (2013) employ Google queries in a sample of the largest European stocks belonging to Euro Stoxx index and at the level of this market index as well. Their results show that the increase in Google queries for stocks precedes the increase in volume, volatility and a drop in cumulative returns. At the level of market index, a negative link is present. An increase in search queries leads to a decrease in the index returns, stock index futures and an increase in implied volatility. Latoeiro et al. (2013) also consider their findings as the affirmative evidence about the presence of retail investors with bounded rationality.

Previous papers focused to Google queries were focused on the trading of less-rational investors in line with ideas of behavioral finance. However, Vozlyublennaia (2014) proposes the reverse argument that higher investors' attention is related to higher market efficiency and not to a higher noise and consequently lower market efficiency as is presented in Da et al. (2011). Vozlyublennaia (2014) argues

with the theory proposed by Grossman and Stiglitz (1980), who suggest that more information or a greater number of informed investors leads to "better" prices because they incorporate more information.

Table III. Key highlights – Google

Reference	Data sample	Transmission mechanism	Market/ Securities	Main Findings
Da et al. (2011)	Standardized (queries)	Information demand	Securities	The increased volume of searches predicts temporary higher stock prices during the next 2 weeks and price reversal follows within the year; Google queries predicts a company's IPO high first-day return and long-term underperformance.
Joseph et al. (2012)	Standardized (queries)	Reaction of a society	Securities	Online search activity reliably predicts abnormal stock returns and trading volumes over a weekly horizon. Sensitivity of returns to search intensity is positively associated with the difficulty of a stock being arbitraged.
Mao et al. (2011)	Standardized (queries)	Reaction of a society	Market	Granger causality between Google queries and market index closing values/volumes. Better forecasting accuracy when Google queries incorporated, especially when the market index is more volatile in the downward trend.
Vlastakis and Merkellos (2012)	Standardized (queries)	Information demand	Securities	Dynamic and contemporaneous link between information demand and supply. Positive association of information demand with historical volatility, implied volatility and trading volume.
Latoeiro et al. (2013)	Standardized (queries)	Information demand	Securities, Market	Negative link between changes in Google queries and stock/index returns; Positive association of searches with volumes and volatility for both, stocks and index.
Vozlyublennaia (2014)	Standardized (queries)	Information demand	Market	Retail investors can create occasional fluctuation but not permanent shifts. Increased attention proxied by Google queries diminishes the predictability of returns and supports the existence of a more efficient market.
Da et al. (2015)	Standardized (queries)	Reaction of a society	Portfolios, Markets	Sentiment index (FEARS) predict return reversals, temporary market volatility and fund flow out of equity funds into bond funds.
Ding and Hou (2015)	Standardized (queries)	Reaction of a society	Securities (liquidity)	Increased investor attention represented by Google queries leads to a reduced relative bid-ask spread and the higher turnover rate.

Vozlyublennaia (2014) finds that presence of attention diminishes the predictability of returns. In other words, it means higher market efficiency. Vozlyublennaia (2014) also argues that retail investors probably do not use Google as a tool for searching for information about individual companies. This argument reflects the reality that retail investors have more options to invest in broader portfolios than in individual stocks due to information and transaction costs. Therefore, Vozlyublennaia (2014) studies the investors' attention represented by Google queries and broad market indices like the Dow Jones Industrial Average, S&P 500, NASDAQ, accounting for large, medium-sized and small companies respectively. Furthermore, she studies investors' attention to gold and oil via the Chicago Board Options Exchange Gold Index and West Texas Intermediate crude oil index, respectively. Vozlyublennaia's (2014) results show affirmative evidence that retail investors can create occasional fluctuation but not permanent shifts. Overall, this evidence supports the existence of a more efficient market due to the increased attention of investors. Vlastakis and Merkellos (2012) studies the information supply and demand and its relationship to the stock market. They argue that information demand can be properly analyzed via Google data as never before. Their argument is based on the paper by Grossman and Stiglitz (1980) as in the study Vozlyublennaia (2014), but they do not literally suggest that higher information demand relates to markets that are more efficient. Still, they provide an interesting analysis focused on 30 of the largest stocks listed on the New York Stock Exchange (NYSE), NASDAQ and on the aggregate market represented by the S&P 500. Their findings show the contemporary and dynamic link between demand and supply. On top of that, information demand is significantly related to historical volatility and trading volume at the level of individual stock, and to the aggregate market as well.

4 Future directions

At this very moment, the employment of social media big data in the field of behavioral finance is at the beginning. All the papers within section 3 can be considered as exploratory studies providing the first insights in this field. This section discusses two areas (challenges) that the author sees as crucial for this field to mature. The methodological challenges rise from the big data processing and the theoretical one deals with the further need to study the behavioral finance theories on a social media data.

4.1 Methodological challenges

Today, as presented above, the dominant part of research employs standardized and publicly available data. Google queries are popular source of data and one can assume that even the Facebook's Gross National Happiness Index would be as well were it still in existence. It is an expected outcome, since

this data is readily available without the employment of sentiment analysis. An advantage of standardized data is its ability to compare the findings across the studies and discuss their robustness. However, a disadvantage, especially when Google queries are employed, is the missing qualitative distinction. This can be a shortcoming when authors consider Google queries as a proxy for sentiment as in the study of Joseph et al. (2012) who employ the search volumes for the stock ticker. Sentiment, by principal is qualitative in nature but the search volume for the stock ticker does not convey one. Therefore, the only lead to evaluate its effect is the performance of regression analysis and the evaluation of the regression coefficient sign. Another solution is the employment of search phrases that convey the sentiment like "bankruptcy" as in study Da et al. (2015). Therefore, one can argue that increasing search volume for bankruptcy is in line with growing negative sentiment in society.

The problem of ambiguous sentiment distinction related to standardized data can be solved via the employment of qualitative data in the form of Facebook comments or Twitter tweets, processed by sentiment analysis. Such data when used in econometric modeling is defined in advance as vectors of positive, neutral or negative values. However, on the contrary, when a unique data sample is employed, it can be difficult to compare the results across the papers. In particular, Mao et al. (2011) imply that the final findings rely on a unique, particular combination of data sets and specific sentiment tracking tools. Similarly, Sprenger et al. (2014a) encourage research on the role of information weighting and dissemination of information in social media because picking the right tweets (potential sample bias) remains just as difficult as making the right trades. These ideas are summarized in Gonçalves et al. (2013) who compare the existing sentiment analysis methods. They argue that current practice is focused on development of applications for social media analysis without a detailed understanding of technicalities and limitations. The main findings of Gonçalves et al. (2013) can be summarized as follows. 1. Coverage problem. Compared sentiment methods cover between 4% to 95% of the sample when employed to real events. It can lead to bias or underrepresentation of data. 2. Polarity problem. Existing methods interpret the sentiment (positive, neutral or negative) in the studied corpus of text differently, ranging from 33% to 80%. Gonçalves et al. (2013) propose a solution for this problem. They provide a publicly available iFeel domain (http://www.ifeel.dcc.ufmg.br) where one can compare the performance of different sentiment analysis techniques. Another solution to increase the robustness of results is the employment of several sentiment sources to describe the economic and financial events as in the study by Mao et al. (2011).

In summary, regarding the methodology, there is a tradeoff between unique and standardized data in terms of a sentiment polarity and representativeness of results. The author considers it as very

important point in the further research, especially a need to define the generally acceptable techniques with robust results for sentiment analysis on social media big data samples. It is a challenging issue because social media big data is often available in the form classified as an unstructured sentiment – an informal and free text format because social media users do not follow any constraints (Arjun et al., 2013, Hussein, 2016). Therefore, the development of sentiment analysis methods that properly evaluate the polarity of sentiment is a difficult and complex computational task.

4.2 Theoretical challenges

In the field of behavioral finance, social media big data is currently predominantly employed for modeling of sentiment in the U.S. market. Therefore, it would be interesting to see the broader discussion about the presence of investor demand and sentiment also on other stock markets. In line with the argumentation of Sayim and Rahman (2015), empirical findings for one country may not be generalizable to other countries due to cultural and societal differences. Further, the above papers are focused especially on individual stock prices or stock indices that represent the market. However, investing is based on a portfolio and its formation at the level of individual investors can be influenced by sentiment and information demand as well. Further, behavioral finance theories also present other interesting topics where behavior of an individual investor is present or sentiment of a society can be influential even for institutional investors. Examples of such topics are initial public offerings, mergers and acquisitions or the effects of social interactions to investing (e.g. herding behavior) where sentiment is present by nature and social media big data can be a unique source of data to provide better insights.

Conclusion

This paper provides an overview in the field of employing social media big data to capital markets. It provides a brief description and functioning of social media, big data and sentiment analysis. The contribution of this paper is in the chapter "*Economic rationale in social media big data applications",* because it describes a logical sequence and the transmission mechanism from social media to capital markets currently missing in the literature. In particular, it emphasizes the role of social media big data only from a technical point of view as a tool tracking the aggregate behavior of the society. However, the economic rationale of social media data applications is based on the economic theory developed in the field of behavioral finance describing the retail investors, formation of their decision and consequent trading. In addition, the paper presents the main findings of several papers and in the section "*Future Directions"*, it outlines the future challenges in this area of research.

References

Annett, M. and Kondrak, G. (2008): Comparison of sentiment analysis techniques: polarizing movie blogs. In: *21st Conference on advances in artificial intelligence*, (2008): 25–35.

Arjun, M., Vivek, V., Bing, L. and Natalie, G. (2013): What yelp fake review filter might be doing. In: *Proceedings of The International AAAI Conference on Weblogs and Social Media* (ICWSM-2013).

Baker, M. and Wurgler, J. (2007): Investor Sentiment in the Stock Market. *National Bureau of Economic Research*. Working Paper 13189: 1-38.

Barber, M. B. and Odean, T. (2008): All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studie*s 21(2): 785–818.

Barber, M. B. and Odean, T. (2011): The Behavior of Individual Investors. Available at SSRN: http://ssrn.com/abstract=1872211.

Barberis, N., Shleifer, A. and Vishny, R. (1998): A model of investor sentiment. *Journal of Financial Economics*, 49(1998): 307-343.

Beyer, A. M. and Laney, D. (2012): *The importance of big data: A definition*. Stamford, CT: Gartner.

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

Boyd, M. D. and Ellison, B. N. (2008): Social Network Sites: Definition, History and Scholarship. *Journal of Computer-Mediated Communication*, 13(1): 210-230.

Bukovina, J. (2015): Sentiment and blue-chip returns. Firm level evidence from a dynamic threshold model. *Mendelu Working Papers in Business and Economics*, No. 2015-53.

Case, E. K. and Shiller, J. R. (2003): Is There a Bubble in the Housing Market? *Brookings Papers on Economic Activity*, 2(2003): 299-362.

Cambria, E., Schuller, B. and Xia, Y. (2013): New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Computer Society*, 2(28): 15-21.

Choi, H. and Varian, H. (2012): Predicting the present with Google Trends. *Economic Record.* 88(2012): 2-9.

Gonçalves, P., Araújo, M., Benevenuto, F. and Cha M. (2013): Comparing and combining sentiment analysis methods. *Proceedings of the first ACM conference on Online social networks.* (2013): 27-38.

Da, Z., Engelberg, J. and Gao, P. (2011): In search of attention. *Journal of Finance*, 66(5): 1461–1499.

Da, Z., Engelberg, J. and Gao, P. (2015): The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies*, 28(1): 1–32.

D'Amuri, F. and Marcucci, J. (2012): The predictive power of Google searches in forecasting unemployment. *Temi di discussion (Economic Working Papers) 891.*

De Long, J. B., Shleifer, A., Summers, L. H. and Waldmann, R. J. (1990): Noise Trader Risk in Financial Markets. *Journal of Political Economy*, 98: 703-738.

Ding, R. and Hou, W. (2015): Retail investor attention and stock liquidity. *Journal of International Financial Markets, Institutions and Money*. 37 (2015): 12-26.

Fang, L. H. and Peress, J. (2009): Media coverage and the cross-section of stock returns. *Journal of Finance* 64(5): 2023–2052.

Fang, X. and Zhan, J. (2015): Sentiment analysis using product review data. *Journal of Big Data.* 2(5):1-14.

Fenzl, T. and Pelzmann, L. (2012): Psychological and social forces behind aggregate financial market behavior. *Journal of Behavioral Finance*, 13(1): 56-65.

Gartner. (2015): Big Data. www.gartner.com/it-glossary/big-data (accessed, November 16, 2015).

Ginsberg, J., Mohebbi, H. M., Patel, S. R., Brammer, L., Smolinski S. M. and Brilliant, L. (2008): Detecting Influenza Epidemics using Search Engine Query Data. *Nature.* 457(7232): 1012-1014.

Godsay, M. (2015): The Process of Sentiment Analysis: A Study. *International Journal of Computer Applications.* 126(7): 26-30.

Grossman, S. J. and Stiglitz, J. E. (1980): On the impossibility of informationally efficient markets. *American Economic Review* 70(1980): 393-408.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005): The predictive power of online chatter. In: *Proceeding of the Eleventh ACM SIGKDD Internation conference on Knowledge Discovery in Data Mining.* (2005): pp.78-87.

Grullon, G., Kanatas, G. and Weston, J. P. (2004): Advertising, breadth of ownership, and liquidity. *Review of Financial Studies* 17(2): 439–461.

Halevi, G. and Moed, H. (2012): The Evolution of Big Data as a Research and Scientific Topic. *Research Trends Special Issue on Big Data* 30: 3-7.

Hussein, M. (2016): A survey on sentiment analysis challenges. *Journal of King Saud University – Engineering Sciences* (2016).

Jansen, J. B., Zhang, M., Sobel, K. and Chowdury, A. (2009): Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology.* 60(11): 2169-2188.

Joseph, K., Wintoki, M. B. and Zhang, Z. (2011): Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting* 27(2011): 1116-1127.

Kahneman, D. and Tversky, A. (1979): Prospect theory: An Analysis of decisions under risk. *Econometrica*, 47(2): 313-327.

Kaplan, M. A. and Haenlein, M. (2010): Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1): 59-68.

Karabulut, Y. (2013): Can Facebook predict stock market activity? Available at SSRN: http://ssrn.com/abstract=1919008.

Kietzmann, H. J., Hermkens, K., McCarthy, P. I. and Silvestre, S. B. (2011): Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3): 241-251.

Kumar, A. and Lee, M. C. C. (2006): Retail Investors Sentiment and Return Comovements. *Journal of Finance* 61(5): 2451-2486.

Lamb, A., Paul, J. M. and Dredze, M. (2013): Separating Fact from Fear: Tracking Flu Infections on Twitter. In: *Conference of the North American Chapter of the Association for Computational Linquistics: Human Language Technologies.* (2013): 789-795.

Latoeiro, P., Ramos, B. S. and Veiga, H. (2013): Predictability of stock market activity using Google search queries. *Statistics and Econometrics Series 05*. Working Paper 13-06.

Liu, B. (2012): *Sentiment Analysis and Opinion Mining*. Morgan and Claypoll Publishers. p. 168.

Mao, H., Counts, S. and Bollen J. (2011): Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. arXiv preprint arXiv:1112.1051, Cornell University Library.

McLaren, N. and Shanbhogue, R. (2011): Using Internet Search Data as Economic Indicators. *Bank of England Quarterly Bulletin* 51(2): 134-140.

Mishne, G. and Rijke, D. M. (2006): Capturing global mood levels using blog posts. In: Nicolov, N., Salvetti, F., Liberman, M. and Martin, H. J. (Eds.), *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.* (2006): 145-152.

Nofsinger, R. J. (2005): Social Mood and Financial Economics. *The Journal of Behavioral Finance*. 6(3): 144-160.

Pang, B. and Lee, L. (2008): Opinion mining and sentiment analysis. *Foundation and Trends in Informational Retrieval*. 2(1-2): 1-135.

Russel, A. M. (2011): *Mining the social web*. 1st ed. Sebastopol, CA: O'Reilly, p. 332.

Sayim, M. and Rahman, H. (2015): An examination of U.S. institutional and individual investor sentiment effect on Turkish stock market. *Global Finance Journal.* 26(2015): 1-17.

Sakaki, M., Okazaki, M. and Matsuo, Y. (2010): Earthquake shakes twitter users: real-time event detection by social sensors. In: *International Conference on World Wide Web* (2010): 851-860.

Saxa, B. (2014): Forecasting Mortgages: Internet search data as a proxy for mortgage credit demand. *CNB Working Paper Series* 14/2014.

Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P. (2014): Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior and Organization,* 107(2014): 730-743.

Shiller, J. R. (1984): Stock prices and Social Dynamics. *Brookings Papers on Economic Activity*. 2(1984): 457-510.

Shleifer, A. and Vishny, W. R. (1997): The limits of arbitrage. *The Journal of Finance*. 52(1): 35-55.

Sprenger, O. T., Tumasjan, A., Sandner, G. P. and Welpe M. I. (2014a): Tweets and trades: the information content of stock microblogs*. European Financial Management*. 20(5): 926-957.

Sprenger, O. T., Sandner, G. P., Tumasjan, A. and Welpe M. I. (2014b): News or noise? Using Twitter to Identify and Understand Company-specific News Flow. *Journal of Business Finance and Accounting*, 41(7): 791-830.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, M. I. (2010): What 140 characters reveal about political sentiment. In: *4th International AAAI Conference on Weblogs and Social Media*.

Vlastakis, N. and Markellos, N. R. (2012): Information demand and stock market volatility. *Journal of Banking and Finance*. 36 (2012) p. 1808-1821.

Vozlyublennaia, N. (2014): Investor attention, index performance, and return predictability. *Journal of Banking and Finance*. 41(2014): 17-35.

Yu, B., Kaufmann, S., Diermeier, D. (2008): Exploring the characteristics of opinion expressions for political opinion classification. In: *Proceedings of the 2008 international Conference on Digital Government Research,* (2008): 82-91.