Big Data Mining for Investor Sentiment

Tao Cen1*, Qianqian Chu1 and Renke He1

1 Business School, Qingdao University, Qingdao, Shandong, 266000, China

*Corresponding author’s e-mail: centao007@126.com

Abstract. Investor sentiment is a key factor affecting asset volatility, but it is hard to quantify. Previously investor sentiment is mainly measured from questionnaires and market exchange data. With rapidly developing of big data, the way of knowledge exploration has been changed. Various types of data are produced and reserved every minute, particularly unstructured Internet big data which can reflect investor behaviour and investor sentiment directly. Investor sentiment will be better quantified through big data. In this paper, we review new data sources and analytic methods for quantifying investor sentiment, and discuss the future of big data in behavioral finance.

1. Introduction
Investor sentiment influences asset price, liquidity and volatility [1]. Although it is important to investors decision-making, investor sentiment is hard to measure. Traditionally, indicators of investor sentiment include subjective and objective indicator. Subjective indicator is based on questionnaires. But due to low survey coverage, it has sample representativeness and reliability problems. Also, survey is very cumbersome and takes a long time. Objective indicator is based on stock exchange data, but it tends to be quite simple, and proxies such as turnover rate and trade volume lack micro-foundation. Therefore, new proxies of investor sentiment should be developed. With rapid development of Internet and Mobile Internet, a large amount of investor related big data is generated online every minute. Big data mining provides new methods for tracking investor behaviour and measuring investor sentiment, and is becoming more and more extensive. Internet textual sentiment is a supplement of traditional sentiment. This paper reviews big data sources and methods for measuring investor sentiment, then summarizes processes of disposing chaos big data, and finally points out advantage and challenges of big data mining for investor sentiment.

2. Understanding big data

2.1. Definition of big data
With development of information technology, the big data era is coming [2]. Yet big data has not a clear definition, it refers to big data set that is difficult to store, read, process and analyse by traditional ways. Big data can be well described by 5V, including: (1) Volume, referring to a large amount of data, in unit of PB, EB or even ZB; (2) Variety, involving to various data types, including structured, semi-structured and unstructured data such as HTML, audio, video, pictures, GPS information and so on; (3) Value, meaning that big data is valuable although relatively low value density; (4) Velocity, relating to high growth speed, and high standard for processing data; (5) Veracity, emphasizing the authenticity and objective record of big data, but different sourced data have different credibility. In a word, big data is big but useful, and has become means of production.
2.2. Big data transform scientific research paradigm
Hidden rules and laws can be better understood with big data analysis. Big data will transform humans’ live, work and think in the following three aspects [3]. First of all, big data use overall data, avoiding drawback of sample survey. Secondly, big data is willing to accept diverse and complex data, and no longer limited to precise data. The more data dimensions, the better. Simple algorithm of big data is more effective than complex algorithm of small data. Finally, big data emphasizes on correlating and predicting, and weakens casualty effect. ‘What is that’ is more important than ‘why is that’. Big data mining bases on massive data induction and statistics, promoting novel relationship exploration and new empiricism formation. Knowledge derives from bottom-up data mining without explicit theory. In the big data era, everything can be quantified [3]. Data can be obtained from the most unlikely places, and then correlations can be found directly through data mining. That has transformed scientific research paradigm to the fourth paradigm — data intensive paradigm. The paradigm can digitize events that were invisible and unmeasurable in the past, and discover new knowledge through big data mining. Therefore, big data has significant impact on the human thinking, cognitive models, social-economic life and value orientation. In particular, Internet big data is a large data sets related to investor sentiment and investor behaviour.

3. Big data sources for mining investor sentiment
Big data is no longer limited to officially published data, because personal and institutional activities are digitalized, recorded and stored. Big data includes structural and unstructured data, the latter includes text, audio, video and image. Because of huge number of Internet users, Internet big data is large volume and continue growing exponentially. Internet textual data has been widely used in sentiment analysis [4]. Because of their relatively easy accessibility and operability, search engine data, social media, stock forum and Internet news have been widely studied for forecasting stock market. We review four main sources as follow.

3.1. Search engine
Investors get information through search engine at anytime and anywhere. Investor attention and investor information demand are directly reflected in searching keywords. Well-known search engines data include Google trend and Baidu Index. Researchers use these indexes extensively to study relationship among investor attention, stock returns, price volatility, market volatility, and market efficiency. Google search volume, as a proxy of investor attention, was used to predicts Russell 3000 Index stocks, showing a positive correlation between the volume and stock prices in the next two weeks, but price reversal eventually within the year. And it also predicts the large first-day return and long-term underperformance of IPO stocks [5]. FEARS index bases on search queries from words like ‘recession’, ‘unemployment’ and ‘bankruptcy’ and so on, and the index predicts return reversals, temporary market volatility and money flow from equity funds to bond funds [6]. Google search volume closely correlate to liquidity of S&P500 stocks as well [7].

3.2. Social media
Social media has become the primary medium for investors to get information. Social media network has two characteristics. First, message and information are spread rapidly on the social network, providing a suitable platform for studying investor sentiment and market response. Investors exchange investment ideas through social media as well. On the one hand, it is profitable to be an influential speaker because people tend to overestimate others’ opinions. Participants in social media are willing to get information from others, especially from the most influential people. On the other hand, once people form their own opinions, they tend to spread these opinions to other individuals. Second, social media records a large number of investors’ emotion, which is useful to study heterogeneity risks and market efficiency. During information spreading in social networks, because of investors having different geography position, investors receive information at different speeds. Therefore, individuals form beliefs or emotions at different speeds. The negative sentiment contained in Twitter messages is
negatively correlate with stock index, and the more negative sentiment will follow with the higher volatility of stock indexes next day [8]. More than 200 million pieces of posts on Twitter about 30 NASDAQ stocks are mining to construct investor sentiment index; this index can predict the stock market movement with an accurate rate up to 70 percent [9].

3.3. **Stock forum**

Stock forum is an online platform for investors to track company news and exchange investment ideas. Investors get public information, forecast information, speculative information, and personal comments from stock forums. Stock forums have some advantages in analysing investor behaviour and sentiment. First, forum posts reflect investor concerns and emotions. Second, posts on stock forum contain disagreements and emotional differences among investors. Third, posts on stock forum contain non-public information, which is useful for predicting stock returns. Finally, the financial online forum topics are rather professional, and posters and readers have knowledge of financial market, thus decreasing noise trading. Textual sentiment, mining from Yahoo! Finance and Raging Bull Message Boards by Naive Bayes algorithm, is positively correlated with stock return. And the post number is negatively correlated with the yield next day [10]. Das and Chen (2007) collect messages from Yahoo Stock Forum, and label each message to bullish, bearish or flat view, then measure small investor sentiment of high-tech stocks, suggesting that the index is closely related to stock volatility [11].

3.4. **Internet news**

Internet news refers to news released by the media on the Internet, including politic, economic, affairs, company dynamics, stocks analysis and so on. Internet news is large number, high timeliness and diversity, and has been used by many scholars to study the relationship between investor sentiment and asset prices. Measuring investor sentiment based on Internet news is indirect. Online news cannot directly represent investor attention. Only when investors receive such information will they pay attention to the securities. Internet news can be divided into objective news and subjective news. Objective news describes the event, and subjective news refers to biased individual report. Therefore, even for a same event, medias publish different news or even the opposite news, which will convey positive or negative emotions to the public. Tetlock et al (2008) built investor sentiment index based on Wall Street Journal stock news, and find that negative sentiment led to stock price fallen [12].

4. **Big data mining process for investor sentiment**

Big data contains quantitative and qualitative information. Quantitative data includes article number, reading number, forwarding number, search volume and collection number and so on, reflecting the demand and attention of investors. Qualitative data directly reflect emotions of market participants, such as article views and commentary information. Several processes for quantifying investor sentiment are as follows. The first step is to get bid data. Using web crawlers, website APIs to get big data sets. The second step is storing and cleaning data. Distributed management system is adopted to ensure standardized access of big data based on data types, formats, update periods. Then, qualitatively and quantitatively analyse text information, for example, marking article with positive, negative, or neutral, then counting article number, reading number and so on. Finally, combining the qualitative and quantitative information to calculate the investor sentiment. Qualitative analysis is the most difficult process and has become the main content of sentiment analysis [13].

Sentiment analysis is mainly based on machine learning and lexicon method. Machine learning can be divided into supervised learning and unsupervised learning. Machine learning classifier algorithm includes K-mean algorithm, maximum entropy, neural network, support vector machine and so on. Lexicon method regards article as any combination of words (i.e., bag of words), ignoring article structure, word order, grammar, and syntax. Then, every word in article is labelled as negative or positive base on predefined dictionary. The highest proportion of the marked mood determines the text mood. The higher proportion of positive vocabulary, the more optimistic article is. Lexicon method face two problems: word coverage and the weight of each word. Higher coverage in the dictionary.
result in more accurate classification. Both machine learning and lexicon method have advantages and disadvantages. Machine learning is easy to use and accurate. But it uses more computing resources, and needs to manually mark the training set. It is necessary to ensure the accuracy of manual marking. So, machine learning relies on the accuracy of manual tagging in some sense. In contrast, the lexicon method does not require manual marking of text, and has a good classification given high vocabulary coverage. However, a vocabulary does not always have the same meaning in different contexts. For example, in the general GI dictionary and Harvard dictionary, 73.8% words are negative, but they are positive or neutral in the financial context [14].

5. Advantages for big data mining on capital market
Measuring investor sentiment and behaviour is hard, as well as mechanism of changing from individual behaviour to overall behaviour. Individual preferences and information disseminations behind Internet big data provide good material for studying investor behaviour and investor psychology. Containing a lot of company valuation information, big data is significant to capital market. Internet big data is large volume and time-sensitive, directly reflecting investor sentiment, and more effective than the previous proxy variables. In addition, Internet big data spread rapidly and interact easily. Owing to the wisdom of crowd, fake news and inferior information are more likely to be eliminated by investors, preserving superior information, and thus market will be more efficient.

Big data is a new type of production material. Investor sentiment mining is just one aspect of big data application in the capital market. Capital market participants, if they have big data resource and data mining technologies, can better understand market movement based on big data mining. Big data mining platform has become an important driving force for institutions competition, expanding business scale and creating value. Big data promotes capital market participation institutions, such as financial institutions, stock exchanges and other mechanisms to create value. Market participant face challenges of how to grasp investment timing and opportunities in a rapidly changing information age. Big data mining improves the reaction speed of investors and is conductive to decision-making. Based on big data, institutional investors can reduce information sharing costs, and improve products and operational efficiency. Big data sentiment analysis may effectively recognize the current market conditions and development trends, improving risk management ability.

6. Challenges and future
Rapid development of information technology provides not only opportunities but also challenges for financial research. First of all, data dimension is quite simple. Many factors affect investor sentiment. There are still many helpful data types such as Online Shopping Platforms, Mobile Payment, GPS Information, Night Light Intensity, and Sensor Data. These data have been used widely in the macro-economic research, but rarely used in researching investor sentiment. More unstructured data are expected to be used. In this context, huge and multi-dimensional big data means that the amount of data is extremely large, and data types and sources are diverse. Because humans’ mood is affected by physical conditions, such as temperature, humidity, air pollutant, noise and so on, Sensor data related to physical condition will be useful to measure investor sentiment. Also, individual sport data along with medical big data should not be neglected. Muti-dimensional data facilitate cross-validation with more robust result. Although big data contains great value, data are not knowledge before algorithms are executed. In asset pricing area, the gap between chaos data and knowledge is big data mining.

Secondly, more suitable algorithms should be developed and applied in quantifying investor sentiment. Machine learning methods have inherent defects such as over-fitting and slow convergence, and rely too much on artificial designing [15]. Market data is complex, traditional artificial neural networks are also difficult to accurately measure investor sentiment, and easy to be influenced. If data quality of Internet big data is low, the model fitting results are hard to be satisfactory. Furthermore, for ambiguous sentences, correct attitude classification by machine learning is less than 30 percent [15].

Last but not least, research paradigm in finance discipline is to be changed. Under traditional research paradigms, it requires so many assumptions that it is difficult to adapt to a rapidly changing
market, resulting in lost a lot of information. And traditional paradigm is not suitable for analysing complex, high-dimensional and high-noise financial data. There is still a gap between financial research and investment application. In contrast, the fourth paradigm is based on big data mining digitalized event and behaviour, and could be mainstream in the near future.

7. Conclusion
Investor sentiment is a factor for asset pricing, but it is difficult to measure. With the widespread use of big data in various fields, data-intensive science has become a new research paradigm, providing new data and method for investor sentiment measurement. The most popular big data source is Internet big data, including textual data from social media, search engine and online forum. These data directly reflect investor sentiment by data mining. But big data mining for investor sentiment still have some flaws, such as low data dimension, and more suitable algorithms are need. And future research paradigm in finance is expected to be changed.

Acknowledgments
This work is supported by the Qingdao Postdoctoral Applied Research Funding (2016044).

References
[1] De Long, J.B., Shleifer, A., Summers, L.H., Robert, J.W. (1993) Positive Feedback Investment Strategies and Destabilizing Rational Speculation. The Journal of Finance., 2: 379-395.
[2] McKinsey Global Institute. (2011) Big data: The next frontier for innovation, competition, and productivity. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
[3] Mayer-Schonberger, V., Cukier, K. (2013) Big data—A Revolution That Will Transform How We Live, Work, And Think. Houghton Mifflin Harcourt Publishing, New York.
[4] Teoh, S.H. (2018) The Promise and Challenges of New Datasets for Accounting Research. Accounting, Organization and Society., 40-69: 109-117.
[5] Da, Z., Engelberg, J., Gao, P.J. (2011) In Search of Attention. The Journal of Finance., 66 (5): 1461-1499.
[6] Da, Z., Engelberg, J., Gao, P.J. (2015) The Sum of All FEARS Investor Sentiment and Asset Prices. The Review of Financial Studies., 28 (1): 1-32.
[7] Ding, R., Hou, W.X. (2015) Retail Investor Attention and Stock Liquidity. Journal of International Financial Markets, Institutions & Money., 37: 12-26.
[8] Bollen, J., Mao, H.N., Zeng, X.J. (2011) Twitter mood predicts the stock market. Journal of Computational Science., 2 (1): 1-8.
[9] Li, B., Chan, K.C.C., Ou, C., Sun, E.F. (2017) Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems., 69: 81-92.
[10] Antweiler, W., Frank, M.X. (2004) Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance., 3: 1259-1294.
[11] Das, S.R., Chen, M.Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science., 53(9): 1375-1388.
[12] Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2007) More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance., 63(3):1437-1467.
[13] Bukovina, J. (2016) Social media big data and capital markets—An overview. Journal of Behavioural and Experimental Finance., 11: 18-26.
[14] Loughran, T., Mcdonald, B. (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance., 66 (1): 35-65.
[15] Su, Z., Lu, M., Li, D. (2017) Deep Learning in Financial Empirical Applications: Dynamics, Contributions and Prospects (in Chinese). Journal of Financial Research., 5:111-126.