ISSN No: 2455–7595

International Journal of Research in Management Studies
A peer reviewed Open access journal.

A STUDY ON BIG DATA MINING FOR INVESTOR SENTIMENT

[1] I R.Kalyani and [2] Swathi Jampala,
[1] Assistant Professor, AMS School of Informatics, OU Road, Hyderabad and
Research Scholar, Department of Business Management, MG University, Nalgonda, India
[2] Assistant Professor, AMS School of Informatics, OU Road, Hyderabad, India

ABSTRACT

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

Keywords: Big data mining, Investor Sentiment, Big data analytics, Investment decision, investment behavior.

Introduction:

Investor sentiment impacts asset price, liquidity, and volatility. Investor sentiment plays an important in investor’s decision-making but it is hard to measure it appropriately. Usually, indicators of investor sentiment include subjective and objective indicators. The subjective indicator is based on questionnaires, but due to low survey coverage, it has reliability problems and a restricted sample. Also, the survey process is very cumbersome and takes a long time. The 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 a micro foundation. Therefore, new proxies of investor sentiment should be developed. With the rapid development of technology, the internet, and mobile internet, a large amount of investor-related big data is generated online every minute. Big data analytics provides different methods for tracking the behavior of investors and measuring investor sentiment. The usage of big data mining is becoming more and more extensive. Internet textual sentiment is a supplement to traditional sentiment. This paper reviews big data mining sources and methods for measuring investor sentiment summarizes processes of disposing of big data and finally points out the advantage and challenges of big data mining for investor sentiment. It also offers the potential to provide investors with an investment decision support mechanism by offering guidelines to help investors and traders determine the correct time to invest in the stock market, stock-yielding returns, etc.,

Objective: This paper examines big data mining sources and methods for measuring and analyzing investor sentiment in investor decision-making.

Research methodology: Secondary data is the main source of inputs. It is taken from various reports available on websites, research articles, and books on big data Analytics. This study is descriptive in nature and attempts to describe the importance of big data analytics in the present Indian Banking scenario.

Definition of big data

With the development of information technology, the big data era is succeeded. "Big

Data" analytics is often used as a synonym for customer analytics, real-time analytics, or predictive analytics. In general Big Data can be defined as a collective term of contemporary methodologies and technologies that collects, organizes, processes, and analyzes large, diverse (structured and unstructured), and multifaceted sets of data. The main characteristics of Big Data Analytics are Volume, Variety, Velocity, and Variability of data.
(1) Volume, referring to a large amount of data, in units of PB, EB, or even ZB;
(2) Variety, involving 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 has different credibility.

Big Data transform scientific research paradigm

Big data analysis is used to understand hidden rules and laws. Big data will transform humans' lives, work, and thinking in the following three aspects. First of all, big data use overall data, avoiding the drawback of sample surveys. Secondly, big data is willing to accept diverse and complex data and is no longer limited to precise data. The more data dimensions, the better. A simple algorithm of big data is more effective than a complex algorithm of small data. Finally, big data emphasizes correlating and predicting and weakens the casualty effect. 'What is that' is more important than 'why that is'. Big data mining is based 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. Data can be obtained from the most unlikely places, and then correlations can be found directly through data mining. That has transformed the scientific research paradigm into the fourth paradigm — data intensive paradigm. The paradigm can digitize actions that are invisible and immeasurable in the past and discover new knowledge through big data mining. Therefore, big data has a significant impact on human thinking, cognitive models, socioeconomic life, and value orientation. In particular, Internet big data is a large data set related to investor sentiment and investor behavior.

Big Data came as a dataset with a size and complexity that is beyond the ability of conventional tools of managing, storing, and analyzing the data. Now, in this context, data also include large amounts of structured, semi-structured and non-structured schemas that can be collected from call logs, social networks, weblogs, emails, and documents. In addition, other sources of unstructured data such as blogs, online news, weather, Twitter, YouTube, and Facebook continuously send out digital information and contribute to what is known as 'Big Data'.

Big Data in Financial Market

The banking and financial institution business involves extensive transactions involving millions daily, each adding another row to the industry's immense and growing ocean of data. To better understand what is forcing Big Data technology adoption in Financial Services, the Oracle white paper titled Information Management and Big Data-A Reference Architecture detailed some drivers that increased the need for Big Data architecture. They are Customer Insight, Regulatory Environment, Explosive Data Growth, and Technology Implication.

The Financial Service sector is one of the most data-driven industries and most of the data that exists within this sector organization's data center that is not analyzed. Financial services organizations are leveraging Big Data mining to transform their processes, their organizations, and soon, the entire industry. Big data is especially promising and differentiating for financial services companies. With no physical products to manufacture data, the source of information became one of their most important assets.

Sources for Big data mining for 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 the huge number of Internet users, Internet big data is a large volume and continue growing exponentially. Internet textual data has been widely used in sentiment analysis. Because of their relatively easy accessibility and operability, search engine data, social media, stock forums, and Internet news have been widely studied for forecasting the stock market. We review four main sources as follows.

- **Search engine**

Investors get information through search engines at any time anywhere. Investor attention and investor information requests are directly reflected in searching keywords. Well-known search engine data include Google Trends and Baidu Index. Researchers use these indexes extensively to study the 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 predpredictsell 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. FEARS index is based 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.

Social media has become the key medium for investors to get information. Social media network has two characteristics. First, messages 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' emotions, which is useful to study heterogeneity risks and market efficiency.

During information spreading in social networks, because investors have different geographic positions, investors receive information at different speeds. Therefore, individuals form beliefs or emotions at different speeds. The negative sentiment contained in Twitter messages is negatively correlated with the stock index, and the more negative sentiment will follow with the higher volatility of stock indexes the next day. More than 200 million pieces of posts on Twitter about 30 NASDAQ stocks are mining to constrain the investor sentiment index; this index can predict the stock market movement with an accurate rate of up to 70 percent.

- **Stock medium**

A 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 analyzing investor behavior and sentiment. First, forum posts reflect investor concerns and emotions. Second, posts on stock forums contain disagreements and emotional differences among investors. Third, posts on stock forums contain non-public information, which is useful for predicting stock returns. Finally, the financial online forum topics are rather professional, and posters and readers know financial Marthe ket, 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 the next day. Messages collected from Yahoo Stock Forum are labeled with a bullish, bearish, or flat view, then measure small investor sentiment of high-tech stocks, suggesting that the index is closely related to stock volatility.

- **Internet news**

Internet news refers to news released by the media on the Internet, including political and economic, affairs, company dynamics, stock analysis, and so on. Internet news is large in number, has 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 reports. Therefore, even for the same event, the media publish different news or even the opposite news, which will convey positive or negative emotions to the public. Tet Lock et al (2008) built an investor sentiment index based on Wall Street Journal stock news, and find that negative sentiment led to stock prices falling.

Process of big data mining for investor sentiment

Big data contains quantitative and qualitative information. Quantitative data includes article number, reading number, forwarding number, search volume, collection number, and so on, reflecting the demand and attention of investors. Qualitative data directly reflect the 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 big data using web crawlers and website APIs to get big data sets. The second step is storing and cleaning data. The distributed management system is adopted to ensure standardized access to big data based on data types, formats, and update periods. Then, qualitatively and quantitatively analyze text information, for example, marking articles with positive, negative, or neutral, then counting article number, reading number, and so on. Lastly, the qualitative and quantitative information is combined to calculate investor sentiment.

Qualitative analysis is the most difficult process and has become the main content of sentiment analysis. It is mainly based on machine learning and the lexicon method. Machine learning can be divided into supervised learning and unsupervised learning. Machine learning classifier algorithm includes the K-mean algorithm, maximum entropy, neural network, support vector machine, and so on. The lexicon method regards an article as any combination of words (i.e., a bag of words), ignoring article structure, word order, grammar, and syntax. Then, every word in the article is labeled as a negative or positive base on a 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 faces two problems: word coverage and the weight of each word. Higher coverage in the dictionary results 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, vocabulary does not always have the same meaning in different contexts. For example, in the general GI dictionary and Harvard dictionary, 73.8% of words are negative, but they are positive or neutral in the financial context

Advantages of big data mining

Measuring investor sentiment and behavior is hard, as well as the mechanism of changing from individual behavior to overall behavior. Individual preferences and information dissemination behind Internet big data provide good material for studying investor behavior and investor phycology. Containing a lot of company valuation information, big data is more effective than the previous proxy variables. In addition, Internet big data spread rapidly and interact easily.

Due to the wisdom of the crowd, fake news and inferior information are more likely to be eliminated by investors, preserving superior information, and thus the 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 resources and data mining technologies, can better understand market movement based on big data mining. Big data mining platform has become a significant 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 participants face challenges of how to grasp investment timing and opportunities in a rapidly changing information age. Big data mining increases the reaction speed of investors and is conducive 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.

Challenges and future

The rapid development of information technology provides not only opportunities but also challenges for financial research. First of all, the 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 macroeconomic 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 conditions will be useful to measure investor sentiment. Also, individual sports data along with medical big data should not be neglected. Muti-dimensional data facilitate cross-validation with more robust results.

Although big data contains a great value, data are not knowledge before algorithms are executed. In the 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. Market data is composite; traditional artificial neural networks are also difficult to accurately measure investor sentiment, and easy to be influenced. If the 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. Last but not least, the research paradigm in the 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 the loss of a lot of information. And traditional paradigm is not suitable for analyzing complex, high-dimensional, and high-noise financial data. There is still a gap between financial research and investment application.

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 methods for investor sentiment measurement. The most popular big data source is Internet big data, including textual data from social media, search engines, and online forums. These data directly reflect investor sentiment through data mining. But big data mining for investor sentiment still has some flaws, such as low data dimension, and more suitable algorithms are needed. And future research paradigm in finance is expected to be changed.

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 the 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] Wang G, Wang T, Wang B, Sambasivan D, Zhang Z, Zheng H, Zhao BY. Crowds on Wall Street: extracting value from social investing platforms, foundations, and Trends in information retrieval. New York: ACM; 2014.

[12] Li G, Zhu H, Cheng G, Thambiratnam K, Chitsaz B, Yu D, Seide F. Context-dependent deep neural networks for audio indexing of real-life data. In: IEEE spoken language technology workshop (SLT). 2012. p. 143–8.