The Applications and Future of Big Data Technology in Financial Investment

Mengqi Liu

School of Software Engineering, Tongji University, Shanghai, China

louisliu@alumni.tongji.edu.cn

Abstract. The application of big data technology to financial investment has been a general trend and research hotspot in recent years. The rapid development of technologies such as artificial intelligence and machine learning has gradually transformed investment activities from manual decision-making to algorithmic decision-making. Big data technology can help investors and enterprises process more information, make more transactions and make more accurate judgments, also bring more profits. However, this technology also has its limitations and problems. This paper introduces the main components of big data technology and its practical application in financial investment, and analyzes the advantages and risks of this technology. Finally, this paper discusses the future development direction of big data technology in financial investment, the attitude of society towards technology development and the issues that investors, enterprises and governments need to pay attention to. This paper also mentions the importance of paying attention to and ensuring data ethics while developing technology. This paper aims to systematically summarize the current status of big data technology, provide suggestions and risk warnings for the use of big data technology in various fields, and provide a reference for the research of big data technology.

Keywords: Big data, data mining, high-frequency trading, stock price prediction, data ethics.

1. Introduction

Financial investment involves a wide range of fields, and the more common transactions include stocks, bonds, funds, futures and other instruments. These financial instruments can be further subdivided. For example, bonds can be divided into sovereign bonds issued by the national financial department or local governments, financial bonds issued by financial institutions, corporate bonds issued by enterprises, and so on. In addition, investment and financing are also an important part of the financial market. In recent decades, venture capital has gradually entered the public eye. Well-known companies such as Intel and Apple were initially supported by venture capital, and the rise of venture capital has also spawned private equity funds, that is, investment funds raised from a small number of institutional investors or individuals through non-public means. Today, private equity funds have become synonymous with high thresholds and high returns.

A few decades ago, entrepreneurs may be able to obtain sufficient interest and financial support from investors with a simple business idea or product prototype. Facing the expansion and maturity of the market, the competition in the field of financial investment is becoming more and more fierce. Today it is difficult for investors to find an opportunity similar to Softbank Group's investment in Alibaba and obtain a return of nearly 3,000 times. Due to fewer investment opportunities and more competition, financial investment has become an increasingly risky field, and it has become much more difficult to make a profit in investment. Facing the huge financial market, manpower can no longer handle all the information generated every day, so people introduce big data technology to help investment.

Almost every investment firm uses data mining to support itself today. Investors use big data technology, artificial intelligence, and machine learning to conduct high-frequency trading and predict future stock prices. Enterprises also use big data technology for auditing and daily management and decision-making. Computers and algorithms have computing efficiency that far exceeds that of human beings, and they can maintain stability and objectivity while processing data at high speed. The use of these technologies has indeed brought huge profits to investors.

However, using big data technology to invest also brings problems. Algorithms are not infallible, and decisions can be made at a loss when they encounter data that the algorithm does not understand. The financial market is complex and changeable, any event may trigger the system to change the investment decision, and for different companies, the optimal solution considered by the algorithm is likely to be approximate, and finally the convergence decisions will affect the volatility and liquidity of the market. In addition, the independence of technology also increases the reliance on the authenticity of data, and the consequences of false data will be disastrous. Additionally, the debate over data ownership, privacy, and security in data collection has been a hot topic in recent years. To solve these problems requires not only further development of technology, but also a clearer understanding of big data technology in society.

In the second part, this paper introduces big data technology and its characteristics, composition and application in various fields. The third part of this paper introduces the application of big data technology in financial investment, including data mining, high-frequency trading, stock price prediction and portfolio optimization and risk management. Then in the fourth part, this paper discusses the advantages and risks of using big data technology in financial investment in the future, and tries to put forward the aspects that need attention in the future development of this technology. Finally, this paper concludes that big data technology still needs continuous improvement to improve its ability to resist financial risks. While using this technology, people also need to recognize the limitations of technology, strengthen management, and not over-rely on it.

2. What is big data

2.1. The Definition of Big Data

Although the term big data has been around for decades, there is still no precise definition. Big data is described by the European Commission as vast quantities of varied data kinds generated from numerous types of sources, including people, machines, and sensors. And big data is defined by the National Science Foundation (NSF) as data that present a challenge to current methodologies because of their quantity, complexity or rate of availability. Although scholars in different fields have different understandings of big data, it is generally believed that big data is a term used to describe data collections that are too massive or complicated for conventional data processing technologies to handle. Different from ordinary data sets, only when the amount of data is large enough can the information be understood or the characteristics of the data can be analyzed, it can be called big data, not just a large amount of unrelated data.

2.2. Characteristics of Big Data

Big data has many characteristics and they are still increasing. The five most common characteristics are known as the "5V's" - Volume, Variety, Velocity, Value, and Veracity.

Volume refers to the amount of data generated and stored, reflecting the size of the data set. With the increase in the scope of data collection and the growing emphasis on data in various industries, the scale of big data has grown from several TB or PB to EB or ZB.

Variety refers to the data types, including structured, semi-structured, and unstructured data. Structured data is usually easy to work with, such as numbers in sequential order. Unstructured data, including text, images and videos, is usually random and requires technical methods to be converted into structured data before it can be used. Semi-structured data is somewhere in between.

Velocity usually refers to the speed of data generation and the speed of data processing. Big data usually generates large amounts of new data constantly, and much of the data is highly time-sensitive, such as data from traffic signals or stock market data. So the speed of data processing must be able to match the speed of data generation, otherwise untimely processing of data may cause problems or losses.

Value refers to the value that can be obtained by processing or analyzing big data, such as analyzing market data to assist decision-making, or analyzing user data to push advertisements more accurately, and promote product sales.

Veracity means that the data must be real and reliable. False or low-quality data will affect the accuracy of the analysis and even lead to wrong analysis results.

2.3. The Main Components of Big Data Technology

1) Data Capture

The first step in big data technology is to obtain data. Generally speaking, the main way to capture data is through APIs that connect computers to third-party systems. The application programming interface is a way for computers to communicate with each other. Through the API, the system can disclose some valuable data without revealing the internal details of the system, thus ensuring the security of the system. Users can also easily obtain the desired data with the help of API, without the need to study the specific details of various systems.

In addition, people often use web crawlers to capture various data on the Internet. Web crawlers can access a large number of websites and archive their content. Some search engines use web crawlers to index a large number of websites to facilitate users to search.

2) Data Storage

After obtaining the data, the next step is to collect the scattered data information and store it in the database. How to store a large amount of data efficiently becomes the key issue in this stage. The traditional enterprise self-built database is limited by the deployment cost and space, and it is difficult to update the hardware and software system in time. In this case, enterprises will face great pressure when faced with the storage demand of large amounts of data. Many practitioners in the financial industry have already turned to cloud storage services. The advantage of cloud services is that when an enterprise encounters greater data writing and storage needs, it only needs to request more storage space from the cloud storage service provider without purchasing any additional hardware devices. In addition, cloud storage can not only ensure data security, but also reduce storage costs and reduce the computing burden on servers.

3) Data Analysis

Data analytics is the key part of big data technology and refers to examining, cleaning, transforming, and modeling data. The purpose of data analysis is to find useful information in large amounts of data and use it to assist research or decision-making. Data analysis usually starts with cleaning up errors and anomalies, ensuring the accuracy of the data, and using technologies such as natural language processing to uniformly transform the data into a structured, easy-to-use type. With the help of mathematical modeling and intelligent algorithms, data analysis can discover regularities and characteristics of data. Finally, after analysis and verification, the obtained results can be used in practical work.

4) Data Visualization

Data visualization is to transform a large amount of quantitative or qualitative data into an easy-to-understand graphical interface, such as tables, charts, statistical graphics, etc. The advantage of data visualization is that it makes the data more intuitive and concise. These visual formats enable users to quickly discover the characteristics of the data, make it easier to understand the meaning of the data and the information contained in it, and help them make decisions.

2.4. Applications of Big Data Technology

Nowadays, all walks of life are widely using big data technology to improve efficiency or assist business. In the field of scientific research, scientists use big data technology for climate simulation, genetic decoding, fluid dynamics calculations, etc. Big data technology not only shortens the time for data processing, but also greatly reduces the cost. In sports competitions, big data technology is used to analyze game data and predict results. Sensors on Formula One cars collect all kinds of data and are used by engineers to adjust the cars. In the medical field, big data technology can help analyze

patient data, assist diagnosis and treatment, and help further research on various diseases. Major media companies also use big data technology to collect user preference data, so as to accurately deliver advertisements and content.

The use of big data technology is also widespread in the financial industry. Big data technology has improved the processing speed of various transactions, and thus quantitative transactions have been born. Investment companies use data analysis technology to establish complex statistical models and use the calculation results as the basis for decision-making. At present, big data technology is applied in various aspects such as investment decision-making, block transactions, and investment portfolio management. As long as the amount of data is large enough, big data technology can demonstrate efficiency beyond traditional manual processing. Therefore, when facing complex and volatile markets such as stocks, bonds, and futures, big data technology has become an indispensable tool for investors. Besides, big data technology can also be used for internal data governance of enterprises, and help enterprises develop better through intelligent auditing and supervision.

3. Applications of big data technology in financial investment

The rapid development of big data technology in recent years has turned the competition in the investment field into a "technical arms race". Investors can use the information obtained by data mining to capture tiny market opportunities and make profits by conducting high-frequency transactions. Using processed data and algorithm models to predict stock prices and optimize portfolio allocation has also become a routine operation for funds and investment companies. In addition, with the assistance of big data technology, investors can not only obtain higher profits, but also reduce risks in investment activities. The following section describes the applications of these several big data technologies in financial investment.

3.1. Data Mining

Data mining is to obtain information from the database and use intelligent methods to process it, discover the unknown useful knowledge and convert it into a form that can be understood and used further. Today is an era of big data, all kinds of data information are scattered in various places, and exist in different forms, some of which may contain important economic value. Therefore, data mining is the foundation of big data technology, and any analysis and research needs good data support. Figure 1 is a common flow of data mining in financial investment.

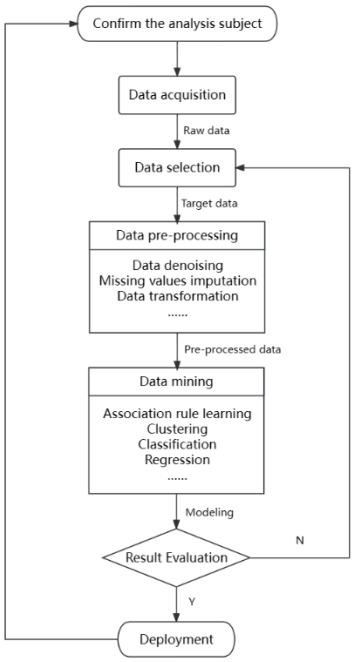

Figure 1. Data mining flowchart. (Figure credit: Original)

1) Data Preprocessing

After determining the analysis theme and selecting the data set, the data information needs to be preprocessed first, including the noise reduction of the data set and the filling of missing values in the data, etc. Noise in data, including invalid data and unexplainable abnormal data, can be removed by models or algorithms. Common models include the linear regression model, Gaussian distribution model, etc. If the original data set can find parameters to fit an approximate linear distribution or approximately conform to a Gaussian distribution, then the data values that deviate from this distribution can be considered as outliers to remove.

However, it is worth noting that abnormal data information is not always noise that should be removed. Sometimes outliers may be caused by human errors, but sometimes outliers are real. For example, among the sales prices of houses in a certain area, the price of a certain house is much higher than that of the surrounding area. It may also be because it is the residence of a celebrity or has other unique values. While outliers not being removed from the dataset can affect the results, valid values being marked as outliers and mistakenly removed can equally corrupt the results.

In addition, when dealing with large-scale data such as stocks, it is very common to encounter data missing, and the way to solve this problem is to fill in the missing value. Commonly used missing value filling methods include mean filling, median filling, mode filling, and before and after data filling, that is, using the mean, median, mode of the data set or the average before and after data of the missing data as the value of the missing data.

2) Common Tasks

In addition to the outlier detection mentioned above, the common tasks of data mining also include association rule learning, clustering, classification and regression. The first two can be summarized as description tasks, and the latter two can be summarized as prediction tasks.

Association rule learning refers to the search for correlations and rules among various types of data or other information carriers. A well-known association rule analysis, known as market basket analysis, states that if a customer buys product A, he is likely to buy product B at the same time, which means that discounting both products at the same time does not significantly increase revenue, and a promotion of one product may boost sales of another. Therefore, by analyzing the correlation between customers' purchased products, it can help companies formulate sales strategies and improve corporate profits.

Two algorithms commonly used in association rule learning are Apriori and FP tree frequency set algorithms. The Apriori algorithm can identify frequently occurring data in the database and gradually expand the data set. As long as the occurrence frequency of the data set is high enough, Apriori can determine the association rules that exist in the database. The FP tree frequency set constructs each transaction into an FP tree, and different projects have different paths in the FP tree. The more paths overlap between two projects, the stronger the correlation between the projects.

Cluster analysis is to group a batch of data or objects so that objects in the same group (also known as a cluster) are more similar and less similar to objects in other clusters. Clustering analysis is easily confused with classification analysis. The difference between the two is that clustering is a process of unsupervised learning, while classification is a process of supervised learning. That is, clustering has no preset categories and completely groups objects according to their similarity. Classification is based on pre-set categories, and objects are assigned to various categories according to similarity. That is to say, clustering first has objects and then categories, and classification has first categories and then objects.

Regression analysis is used to estimate the relationship between label data and feature data, such as the linear regression mentioned above. Support vector machine (SVM) in machine learning is often used for data classification and regression analysis. SVM maps training data to points in space, maximizes the difference interval between two types, and finds the optimal decision boundary. By mapping new data into the same space. SVM predicts the type of data based on which side it falls on, thus assessing whether there is a linear relationship between the data.

After data mining, the processed data can be used for various analyses. For example, Kim et al. mentioned a method of using association analysis to analyze the correlation between keywords on the Internet and discover new investment opportunities. In addition, data mining can also provide support for subsequent trading activities, such as high-frequency trading and stock price prediction.

3.2. High-frequency Trading

As a type of quantitative trading, high-frequency trading is the behavior of using market spreads, price fluctuations and other reliable parameters found in information mining to conduct frequent transactions through preset algorithms and trading strategies. It is characterized by high speed, high turnover rate and high transaction frequency [1].

With the development of machine learning and data analysis, quantitative trading conducted by computers has completely changed the way markets trade. With well-written algorithms, computers can analyze large amounts of data at the same time, and while it is still not possible to program all the elements involved in human analysis, the need for rapid decision-making is clearly increasing in fast-changing financial markets.

With the continuous development of technology, the execution time of high-frequency trading has grown from a few seconds at the beginning of the century to the microsecond or even nanosecond level today [2]. Because high-frequency trading focuses on short-term investments in the market, the profit margin is usually not high. But thanks to the extremely high transaction frequency, high-frequency trading companies can make up for the low profit margin with more than one million transactions.

High-frequency trading typically makes money by exploiting the momentary pricing inefficiencies of actively traded commodities, such as the spread of the same stock on different exchanges, or the small movements in the share price over a given period [1]. In order to reduce transaction latency, high-frequency trading companies usually ensure that they can receive data directly from stock exchanges or other markets, known as direct market access (DMA), and minimize the physical delay between the company's server and the exchange. To reduce data transmission delays, several companies placed their servers next to the New York Stock Exchange.

Several authors have argued in their papers that at the core of high-frequency trading are the algorithms that control trading strategies. But in fact, the top priority of high-frequency trading is the speed of the transaction, rather than trying to use more complex algorithms for data processing. Because for high-frequency trading, it is difficult to quantify the improvement of trading profits by optimization algorithms, but the resulting reduction in data processing speed is very likely to lead to the loss of investment opportunities or opportunities being preempted by other investors. Therefore, the efficiency and accuracy of information acquisition and data mining are particularly important.

Although high-frequency trading uses quantitative models to make investment decisions based on data information, maintains relatively stable returns and relatively low risks with high trading volume and low profit margins, high-frequency trading does not never lose money. Compared with non-high-frequency trading, high-frequency trading has a significantly lower diversity of trading strategies and is more sensitive to changes in market liquidity.

The pursuit of speed in high-frequency trading leads to algorithm models that are usually relatively simple arbitrage algorithms, and investors who follow similar strategies may make similar investment decisions, and eventually the accumulation of a large amount of investment in a short period of time leads to abnormal price changes. Flash Crash on May 6, 2010 is widely believed to be caused by high-frequency trading. On May 6, 2010, the three major US stock indexes all showed a downward trend due to worries about the Greek debt crisis. But around 2:32 p.m., the Dow Jones Industrial Average plummeted nearly 1,000 points, its second-largest intraday drop ever. Afterward, the regulator believed that high-frequency traders tended to exit the market in the face of uncertainty, and the stock market fluctuated greatly that day, so high-frequency traders actively sold, resulting in a surge in trading volume and exacerbating the decline in the index. Procter & Gamble's inadvertently

large sell order triggered massive algorithmic trading as numerous high-frequency traders used similar algorithms, leading them to make the same sell-off.

At present, some authors have proposed new high-frequency trading algorithms based on neural networks or dynamic programming, such as the model based on recurrent neural networks by Cao et al., with the help of cutting-edge technologies such as machine learning, compared with traditional models, it can obtain higher investment returns while maintaining lower risks [3]. In the future, how to ensure the speed of decision-making in high-frequency trading while avoiding similarity, which will lead to follow-up speculation and panic selling, will be one of the focuses of research.

3.3. Stock Price Prediction and Portfolio Optimization

Stock price prediction refers to trying to determine the changing trend of the stock market and the future value of the stock, and profit from it [4]. When big data technology is utilized, stock price prediction is the most common in the field of financial investment. People usually want to be able to predict the future market trend, so as to ensure that their investment decisions are accurate. In practical applications, investors use the mined data (market information, historical returns, investor sentiment, etc.) and machine learning to build models to predict future stock prices, and according to the forecast results, adjust and optimize the stock selection and weight in the portfolio to pursue higher returns.

In some markets, such as China's A-share market, it is difficult for investors to obtain stable returns through high-frequency trading, because there are some restrictions on transactions, such as not being able to sell shares the same day they are bought, and high transaction fees. Therefore, if the profit margin of the transaction is relatively low, it may even be a loss after deducting the transaction fee. In this case, investors need to predict the price trend in the short and medium term, so as to guide investment behavior.

Markowitz, a famous economist and Nobel laureate in economics, once proposed the classic mean-variance model, which introduced mathematical statistics into investment portfolios for the first time. He recommended that investors diversify their holdings as much as they can to lower investment risks [5]. However, if the investment is too diversified, the investment cost that investors need to pay will also increase, which will reduce the expected return of the investment portfolio.

Zhou et al. proposed a portfolio optimization model based on data analysis and machine learning, which evaluates the investment value of companies by analyzing historical stock transaction data, financial indicators of companies, and related news reports. After that, they used the support vector machine (SVM) to predict the future stock price trend, adjusted the stock selection plan, and completed the construction of the investment portfolio. The accuracy and precision of the model for 182 stock price predictions reached above 0.6 [6].

In addition, other researchers have applied various machine learning techniques to stock price prediction. Table I lists the proposed time of some models and the accuracy of tests [4, 7, 8].

Table 1. Time, technology and accuracy of stock price prediction models

Time	***Technology***	***Accuracy***
2006	decision tree	82%
2008	support vector machine (SVM)	61.73%
2011	support vector machine (SVM)	96.46%
2011	Bayesian network	92%
2012	decision tree	65.41%
2013	neural network	59.38%
2013	neural network	87.50%
2015	Bayesian network	76%
2015	Bayesian network	86%
2018	Long Short-Term Memory (LSTM)	62.87%
2019	Attention-based LSTM with sentiment analysis	54.58%
2020	support vector machine (SVM)	75.48%
2020	Logistic Regression	89.45%

Some prediction models show high prediction accuracy, but there are still some differences between data simulation and real transactions, and the evaluation criteria of different studies are also different, so it is difficult to judge the effect of the model by accuracy alone. Yin et al. propose a new predictive model that can incorporate more dynamic factors into the analysis than traditional models. They use Long Short-Term Memory (LSTM) to improve the model. The traditional Recurrent Neural Network (RNN) can only process certain short-term information, and it is easy to forget relatively long-term information, so the more recent information has a greater impact on the model. LSTM adds filtering of past information, so that LSTM has a better performance in long-term data information, which also happens to meet the needs of predicting stock price trends. They used this model to design an automatic trading system, put it into real fund product investment, and obtained an annualized rate of return of 44.71% within three months, and the model worked well [7].

In addition, investors' opinions on investment expressed on social media or in daily life will influence other investors and change their investment willingness. Therefore, investor sentiment will affect the rise and fall of corporate stock prices [9]. Some news related to the company may also have an impact on investors because of the positive or negative information conveyed, which in turn affects the stock price. Research shows that compared with mature markets, Growth Enterprises Market (GEM) is more susceptible to factors such as investor sentiment and industrial changes [4]. Some researchers use RNN and LSTM to analyze investor sentiment and economic news headlines and predict stock prices. Although the results have certain limitations, they can assist investors in making decisions [9, 10].

In actual investment activities, due to the randomness of the financial market, the investment portfolio generated based on historical data may not be optimal, and even the price prediction may not conform to the future trend. Therefore, researchers are also exploring new models and technologies to reduce the impact of market fluctuations on prediction, and further improve the comprehensiveness of analytical data to improve the model's anti-risk ability and improve the profitability of investment portfolios.

3.4. Risk Management

Market risk is the possibility of suffering losses as a result of fluctuating market prices or other market-related events, such as changes in interest rates, equity prices, foreign exchange rates, commodity prices, and other factors [11]. There are also credit and liquidity risks in the banking and insurance sectors. Risk management is to identify and evaluate risks and reduce risks or losses caused by risks in various ways.

1) Enterprise Risk Management

For an enterprise, a temporary loss in investment activities can hardly cause the enterprise to fall into trouble, but a mistake in decision-making or a risk in operation may be fatal. Therefore, risk management is one of the keys to an enterprise's financial activities.

Due to the inability of the traditional risk control model to conduct an in-depth analysis of the data, the evaluation is not comprehensive enough, and the enterprise's anti-risk ability is low [12]. With the development of big data technology, when enterprises make investments, the quantitative risk assessment system established by data mining technology can make decision-making more scientific and reduce the error rate and blindness of investment [13]. The use of machine learning and neural networks to build models can systematically evaluate the feasibility, expected benefits, and risks of investment projects in terms of market, finance, and management, and help enterprises make investment decisions [14].

The audit work of enterprises needs to find problems in a large amount of unstructured data [15]. Today's data centers can use big data audits to continuously collect various types of data from enterprises, analyze business data through artificial intelligence algorithms, discover risk points in business operations, and summarize unstructured data into visual reports to facilitate managers to grasp the business situations.

2) Banking and Insurance Risk Management

Nowadays, one of the main risks faced by commercial banks and insurance companies is credit risk, credit default and insurance fraud have caused a lot of economic losses. Through data mining, commercial banks can collect all kinds of information about customers, including financial status, credit rating and stability of future development, and generate customer portraits. Banks can then use technologies such as neural networks to analyze the default risk of each user, target each customer with precision marketing and set up personalized products [15].

Similarly, banks can conduct risk analysis on enterprises applying for financing, and use data analysis to assess the financing risks of enterprises through tax data, bank statements, corporate financial statements, senior executives' personal information and other data [16]. Zhou et al. proposed a credit risk assessment algorithm based on the BP neural network, which realized the identification of enterprise financial risk. Meanwhile, by storing data distributed on cloud servers, the time required for model training was greatly shortened [17]. With the increasing amount of data, the accuracy of the risk assessment model will also be improved. Banks can quantify the credit limit of enterprises according to various business information of enterprises, so as to better support the development of enterprises and reduce default behaviors.

The insurance industry is similar to the banking industry. Insurance companies can identify the risk of customer service insurance fraud through the analysis of various customer data, and sell precisely customized products to customers. The customized content includes the insurance amount, insurance period, and even a certain deductible clause. Insurance companies can also understand the sales of various insurance products through the analysis of their own business data, so as to adjust the company's business strategy and direction.

4. The future of big data technology in financial investment

With the popularization of big data technology in financial investment, the advantages of technology finance are gradually recognized by the public. However, while bringing a lot of wealth to many investors, the development of big data technology has also brought new problems that need to be solved. What direction will big data technology develop, what attitude should various industries take toward this rapidly developing technology, and what the future of big data technology will be in the financial industry are all topics worth thinking about.

4.1. For Enterprise

Investment and financing is one of the key activities for enterprise development. In the future, big data technology will be more involved in enterprise financing activities, using data mining and data analysis to establish credit risk assessment models and optimize the efficiency of enterprise financing. In addition, when the enterprise invests externally, big data technology can analyze the operating conditions and development prospects of the investment object through a large amount of data, assisting the enterprise in making investment decisions and reducing investment risks.

4.2. For Investor

1) Opportunities in the Era of Big Data

In the past, when investors chose their investment targets, they often needed many professionals to spend a lot of time studying market information, business operations, and industry development, and ultimately made investment decisions through the conclusions of manual analysis, which was not only inefficient, but also inaccurate.

Nowadays, big data technology can reduce investors' learning costs and provide more analysis angles and data support. Generally speaking, each department of an investment bank is only responsible for a single area of investment business, because the knowledge barriers in each area make it very difficult to identify quality investment targets, especially those involving new technologies or businesses with strong specialization. An expert in artificial intelligence, for example, may have no up-to-date knowledge of the pharmaceutical industry. Big data technology, however,

lowers the threshold of industry analysis. After model calculation and systematic analysis, various indicators of different industries are summarized into a similar format, and analysts with sufficient investment experience and financial knowledge can understand the operating conditions of companies in many industries, thus making investment decisions quickly and efficiently.

Nowadays is an era of information explosion, people's lives are filled with all kinds of information and they may miss a lot of critical information. The data mining system will continuously collect all kinds of information in the market and summarize it into visual reports to help investors catch the changes in the market in time so as to discover new investment opportunities.

2) Limitations of Technology

Admittedly, big data technology has made a great contribution to the investment industry, and countless examples have proved that decisions made by AI are not only more efficient than those made by human beings, but even have a higher rate of return. In the past two years, the idea of artificial intelligence replacing humans has become more and more popular. However, in the foreseeable future, big data and artificial intelligence still cannot replace practitioners in the financial investment industry. Existing technology is not yet able to effectively determine the systemic risk of the financial market, when there is a black swan event like COVID-19, many key decisions still need to be made by professionals.

In addition, as mentioned above, analytical models based on the same theory may make similar investment decisions, resulting in lower profits or even losses. For example, in the case of Gamestop stock in January 2021, many large Wall Street investors generally believed that the company's stock price was inflated and shorted it. However, the unanimous buying of many retail investors pushed the stock price to a new high.

This is not just a protest against Wall Street, but also a high-risk speculative transaction by retail investors [18]. This incident does not mean that retail investors who lack the support of analytical data and intelligent algorithms can easily defeat the senior investment banks on Wall Street. In fact, the company's operating conditions did not match its stock price at the time, but this incident also shows that big data technology cannot help in certain complex situations.

Especially when all the analytics point to a business opportunity or investment decision, it may not be a good one anymore. Being different is crucial in investing, and investors can only make significant profits by identifying opportunities that others have not yet realized. Obviously, it is still impossible for a computer algorithm to achieve this.

What's more, investors still need to pay attention to the authenticity of the data, because the premise of data analysis is that all data is real, and false data will lead to wrong analysis results. The current data technology is not yet able to identify all audit frauds of enterprises. In fact, new technologies can be used not only to identify data fraud, but also to data fraud [19]. So this will be a long-standing problem, and the counterfeiting incidents of Enron in the United States and Wirecard in Germany have illustrated this point.

4.3. For Government

1) Promoting the Healthy Development of Financial Market

Just like the rise of electronic stock trading in the 1980s, the application of artificial intelligence, machine learning and other technologies in financial investment has had a huge impact on this industry, but it cannot change its essence. The financial industry has always been a game between people. For the government, how to make the regulation of the industry keep up with the development of technology will be a key issue in the age of data.

As mentioned above, Wirecard’s fraudulent behavior reflects that the current financial market regulatory authorities do not have sufficient regulatory capabilities, and the asymmetry of technical levels between different regions and companies also provides convenience for this fraudulent behavior. Facing the advent of the era of technological finance, both audit firms and government regulators must accelerate transformation to match the development speed of financial companies, and restrict new types of data fraud through new regulatory technologies and policies. Big data

technology should be used in a direction that is beneficial to the development of society and industry, rather than technological monopoly or vicious competition.

2) Focusing on Data Ethics

Data ethics is to require the government, companies or other data users to use data in accordance with ethical and social requirements. The huge amount of data and computing power have created countless opportunities and wealth for many companies, but as the collection and use of data gradually permeates all aspects of life, such as smart recommendations on social media and shopping sites, the issues of data ownership, security, and privacy have caused widespread concern and worry in society. Advances in technology have made data management more difficult instead [20]. People are worried that the collection and analysis of data by large companies will violate their own privacy. The opaqueness of data and processing methods in companies makes it impossible for people to understand what their data is being used for. Whether artificial intelligence algorithms can lead to bias and systematic oppression of specific groups due to data differences, and who should own the value and benefits generated by using personal data are currently highly debated topics.

As a government, it can restrict the use of data by enterprises through legislation and taxation. Penalties for violations of data privacy can limit corporate data abuse and maintain data ethics [21]. In addition, using artificial intelligence and big data analysis technology to deal with data security issues is also a major research direction at present. But this may also raise new data ethics issues. Only by promoting the whole society's attention to data ethics and making consumers at the lowest end of the data industry chain realize the importance of data security and privacy, can data ethics become a socially recognized rule.

However, data ethics remains a very complex topic. For example, sharing medical data can help the research of diseases and promote the development of medicine. This kind of data sharing is obviously very valuable to society, but it contradicts the privacy of personal data. How to measure the value of data and how to protect the freedom and privacy of data on a reasonable scale are issues that the government needs to consider. In addition, data disclosure or protection led by the government may have a great impact on certain industries, and some consequences are unforeseeable [22]. After formulating data management methods and laws, the government's continuous governance will be the key to ensuring data ethics.

5. Conclusion

This paper introduces the application of big data technology in financial investment, including data mining, high-frequency trading, stock price prediction and portfolio optimization and risk management. Then this paper discusses the advantages and risks of using big data technology in financial investment in the future. Big data technology can improve the investment efficiency of enterprises and investors and help them make more reasonable decisions, but big data technology also has defects that cannot resist systemic risks. For example, the nature of high-frequency trading limits the complexity of algorithms, making it more prone to miscalculation when systemic risk arises than investment based on fundamental analysis. Similar problems exist in stock price prediction. When a "black swan event" occurs, existing technologies and algorithms still cannot identify the risks involved, resulting in inaccurate results and even system failure. In addition, the system's ability to identify false data is still very low.

This paper attempts to propose the improvement direction of this technology in the future development, including further research and improvement of the algorithm to improve the anti-risk ability, using big data technology to identify data fraud, promoting the benign development of financial markets, improving social attention to data ethics and strengthening supervision.

Finally, this paper concludes that in today's information age, using big data technology to process data has obvious advantages in efficiency, and the use of big data technology will become more and more common in the financial field in the future. However, people should not only learn to use big data technology, but also recognize the risks and problems brought by technology and constantly

improve it. This paper summarizes the advantages and disadvantages of big data technology, and puts forward the precautions and suggestions on big data technology, which helps to enhance the understanding and risk awareness of this technology. This paper can also serve as a reference for researchers on the applications of big data technology in the field of financial investment, and proposes several future research directions.

The study of this paper also has some shortcomings. There are many applications of big data technology in the field of financial investment, and this paper only selects some of them for analysis. Additionally, in the face of the violent market fluctuations during the epidemic, traders cancelled orders and even suspended trading, resulting in less research data and certain limitations in the analysis conclusions. Future studies can analyze more areas and refer to more data.

References

[1] X. Jia and R. Y. K. Lau, "The control strategies for high frequency algorithmic trading," in 2018 IEEE 4th International conference on control science and systems engineering (ICCSSE), 2018: IEEE, pp. 49-52.

[2] M. Aquilina, E. Budish, and P. O'neill, "Quantifying the high-frequency trading "arms race"," The Quarterly Journal of Economics, vol. 137, no. 1, pp. 493-564, 2022.

[3] X. Cao et al., "A novel recurrent neural network based online portfolio analysis for high frequency trading," Expert Systems with Applications, vol. 233, p. 120934, 2023.

[4] P. Yu and X. Yan, "Stock price prediction based on deep neural networks," Neural Computing and Applications, vol. 32, pp. 1609-1628, 2020.

[5] H. M. Markowits, "Portfolio selection," Journal of finance, vol. 7, no. 1, pp. 71-91, 1952.

[6] Z. Zhou, M. Gao, H. Xiao, R. Wang, and W. Liu, "Big data and portfolio optimization: a novel approach integrating DEA with multiple data sources," Omega, vol. 104, p. 102479, 2021.

[7] T. Yin, C. Liu, F. Ding, Z. Feng, B. Yuan, and N. Zhang, "Graph-based stock correlation and prediction for high-frequency trading systems," Pattern Recognition, vol. 122, p. 108209, 2022.

[8] P. Soni, Y. Tewari, and D. Krishnan, "Machine Learning approaches in stock price prediction: A systematic review," in Journal of Physics: Conference Series, 2022, vol. 2161, no. 1: IOP Publishing, p. 012065.

[9] S. Wu, Y. Liu, Z. Zou, and T.-H. Weng, "S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis," Connection Science, vol. 34, no. 1, pp. 44-62, 2022.

[10] L. Nemes and A. Kiss, "Prediction of stock values changes using sentiment analysis of stock news headlines," Journal of Information and Telecommunication, vol. 5, no. 3, pp. 375-394, 2021.

[11] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," Risks, vol. 7, no. 1, p. 29, 2019.

[12] Y. Song and R. Wu, "The impact of financial enterprises' excessive financialization risk assessment for risk control based on data mining and machine learning," Computational Economics, vol. 60, no. 4, pp. 1245-1267, 2022.

[13] B. Cui, "Research on Big Data Risk Control Model of Venture Capital," in Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing, 2021, pp. 173-181.

[14] L. Deng and Y. Chang, "Risk Management of Investment Projects Based on Artificial Neural Network," Wireless Communications and Mobile Computing, vol. 2022, 2022.

[15] W. Peng, "Research on the Application of Big Data in Financial Investment Risk Management and Control," in 2022 3rd International Conference on Big Data and Informatization Education (ICBDIE 2022), 2022: Atlantis Press, pp. 1139-1147.

[16] J. Li, "Venture financing risk assessment and risk control algorithm for small and medium-sized enterprises in the era of big data," Journal of Intelligent Systems, vol. 31, no. 1, pp. 611-622, 2022.

[17] H. Zhou, G. Sun, S. Fu, J. Liu, X. Zhou, and J. Zhou, "A big data mining approach of PSO-based BP neural network for financial risk management with IoT," IEEE Access, vol. 7, pp. 154035-154043, 2019.

[18] T. Hasso, D. Müller, M. Pelster, and S. Warkulat, "Who participated in the GameStop frenzy? Evidence from brokerage accounts," Finance Research Letters, vol. 45, p. 102140, 2022.

[19] S. Zeranski and I. E. Sancak, "Does the 'Wirecard AG' Case Address FinTech Crises?" Available at SSRN 3666939, 2020.

[20] A. Yarali, R. Joyce, and B. Dixon, "Ethics of big data: privacy, security and trust," in 2020 Wireless Telecommunications Symposium (WTS), 2020: IEEE, pp. 1-7.

[21] J. M. Puaschunder, "Big data ethics," Puaschunder, JM (2019). Journal of Applied Research in the Digital Economy, vol. 1, pp. 55-75, 2019.

[22] B. C. Stahl and D. Wright, "Ethics and privacy in AI and big data: Implementing responsible research and innovation," IEEE Security & Privacy, vol. 16, no. 3, pp. 26-33, 2018.