Literature Review On
“Stock Market Prediction and Analysis using Hadoop”

Aayush Gaigowal
Student, Institute of Industrial and Computer Management and Research (IICMR), Savitribai Phule Pune University
aayushgaigowal1961@gmail.com

Abhishek Behera
Student, Institute of Industrial and Computer Management and Research (IICMR), Savitribai Phule Pune University
abhishekbehera19901@gmail.com

Aachal pardeshi
Student, Institute of Industrial and Computer Management and Research (IICMR), Savitribai Phule Pune University
achalpardeshi159@gmail.com

Literature Review

***Abstract*— This Stock market prediction and analysis play a crucial role in assisting stakeholders with investment decisions and maximizing profits. With the advent of big data, the volume and complexity of financial data have increased exponentially, posing challenges for accurate and efficient prediction models. This research paper presents a novel approach to stock market prediction and analysis by leveraging the power of Hadoop, a distributed computing framework.**

The proposed system integrates Hadoop with the Naïve Bayes algorithm to harness the advantages of parallel processing and data mining techniques. The architecture encompasses various components, including user registration, data upload, prediction analysis, and automation for purchase or sale. By utilizing the Naïve Bayes algorithm, the system considers factors such as previous data, mean calculation, and classification to predict stock market trends.

The key benefit of employing Hadoop in this context is its ability to process and analyse vast amounts of financial data efficiently. By distributing the processing tasks across multiple nodes, Hadoop enables faster access to large datasets and enhances the accuracy of predictions. Moreover, the system incorporates the Map Reduce programming model to handle the complexities associated with big data processing.

The experimental evaluation of the proposed system demonstrates promising results, achieving accuracy rates of 70-80% and significantly reducing processing time compared to traditional approaches. Inexperienced investors can rely on the system's output to make informed investment decisions, thereby mitigating risks and improving profitability.

Overall, this research paper showcases the potential of utilizing Hadoop and the Naïve Bayes algorithm in stock market prediction and analysis. The proposed system offers a transparent and automated approach, empowering stakeholders with reliable insights and enhancing the efficiency of investment strategies in the dynamic and competitive stock market environment.

***Keywords*— Stock market prediction, Hadoop, Naïve Bayes algorithm, Big data processing, Data mining**

I. INTRODUCTION

1.1 Background of stock market prediction and analysis:

Stock market prediction and analysis play a crucial role in financial decision-making for investors, traders, and financial institutions. The ability to accurately predict stock market trends, such as the rise or fall of stock prices, can lead to significant financial gains. However, the stock market is highly volatile and influenced by various factors, including economic indicators, company performance, geopolitical events, and investor sentiment. Predicting these trends is challenging due to the complexity and dynamic nature of the market.

Historically, stock market prediction relied on traditional approaches such as technical analysis, which involves studying price charts and indicators, and fundamental analysis, which evaluates a company's financial health and market position. While these methods provide valuable insights, they have limitations in capturing complex patterns and adapting to changing market conditions.

1.2 Significance of using artificial neural networks (ANN):

Artificial neural networks (ANN) have gained significant attention in stock market prediction due to their ability to analyse and learn from complex datasets. ANNs are computational models inspired by the structure and function of the human brain, consisting of interconnected nodes or neurons. These networks can recognize and process patterns, making them suitable for analysing stock market data.

The significance of using ANN in stock market prediction lies in their capability to handle nonlinearity and capture intricate relationships among multiple variables. Unlike traditional approaches, ANNs can identify hidden patterns and nonlinear dependencies that may impact stock prices. They can analyse large volumes of historical data, including financial indicators, news sentiment, and market trends, to identify patterns and make predictions.

Moreover, ANNs have the advantage of adaptability and learning from new information. They can update their internal weights and connections based on feedback and new data, allowing them to adjust their predictions as market conditions change. This adaptability is particularly valuable in the dynamic and evolving stock market environment.

By leveraging the computational power and pattern recognition capabilities of ANNs, stock market prediction models can potentially improve accuracy and provide valuable insights for investors and financial institutions. However, it is important to understand the limitations and challenges associated with ANN-based predictions, such as overfitting, interpretability, and data quality issues..

II. SUMMARY HADOOP , NAIVE BAYES, BAYES THEOREM

A. Hadoop:

Hadoop is a distributed computing framework designed to handle and process large volumes of data across clusters of computers. It provides a reliable, scalable, and fault-tolerant environment for storing and processing big data. At the core of Hadoop is the Hadoop Distributed File System (HDFS), which stores data across multiple machines in a distributed manner. HDFS divides data into blocks and replicates them to ensure data availability and reliability.

Hadoop utilizes the MapReduce programming model to process and analyze data in parallel. MapReduce breaks down tasks into smaller sub-tasks, distributes them across the cluster, and then aggregates the results. This parallel processing approach enables faster data processing and analysis, making Hadoop well-suited for handling big data applications.

Hadoop offers several advantages, including scalability, as it allows for easy expansion of computing resources by adding more machines to the cluster. It also provides fault tolerance, as it replicates data across multiple nodes, ensuring data availability even if some nodes fail. Hadoop's distributed processing capability enables efficient data processing by utilizing the computational power of multiple machines simultaneously.

B. Naïve Bayes Algorithm:

The Naïve Bayes algorithm is a probabilistic classification algorithm commonly used in machine learning and data mining tasks. It is based on Bayes' theorem, which provides a mathematical framework for calculating conditional probabilities. The algorithm assumes that all features in the data are independent of each other, hence the term "naïve." Although this assumption is often violated in real-world scenarios, Naïve Bayes still produces robust and efficient results.

The Naïve Bayes algorithm calculates the probability of a hypothesis given the observed data by multiplying the prior probability of the hypothesis by the likelihood of the data given the hypothesis. It then normalizes the probabilities to obtain a probability distribution over the possible hypotheses. This distribution is used for classification, where the hypothesis with the highest probability is selected as the predicted class.

One of the key advantages of the Naïve Bayes algorithm is its simplicity and computational efficiency. It requires a relatively small amount of training data and can handle high-dimensional feature spaces. Naïve Bayes performs well in many real-world applications, such as text classification, spam filtering, and sentiment analysis.

C. Bayes Theorem:

Bayes' theorem is a fundamental concept in probability theory named after Reverend Thomas Bayes. It provides a way to update prior probabilities based on new evidence. The theorem states that the posterior probability of a hypothesis given the observed data is proportional to the likelihood of the data given the hypothesis multiplied by the prior probability of the hypothesis.

Mathematically, Bayes' theorem can be expressed as:

$P(h|D) = P(D|h) * P(h) / P(D)$

where $P(h|D)$ is the posterior probability of hypothesis h given data D, $P(D|h)$ is the likelihood of data D given hypothesis h, $P(h)$ is the prior probability of hypothesis h, and $P(D)$ is the probability of the observed data.

Bayes' theorem allows for the integration of prior knowledge or beliefs with observed evidence to obtain updated probabilities. It provides a principled approach for reasoning under uncertainty. In the context of the Naïve Bayes algorithm, Bayes' theorem is used to calculate the probabilities of different classes given the observed features, enabling classification based on the highest probability.

Bayes' theorem has wide-ranging applications in various fields, including statistics, machine learning, and artificial intelligence. It provides a foundation for probabilistic reasoning and inference, allowing for the incorporation of prior knowledge and iterative updating of beliefs based on new evidence.

III. UNDERSTANDING OF SYSTEM ARCHITECTURE

The system architecture mentioned in the paper proposes a framework for stock market prediction and analysis using Hadoop and Naïve Bayes algorithm. The architecture consists of several key components:

1. User Registration: Users are required to register by providing their personal details to access the system.

2. Admin Authentication: The system includes an admin component that authenticates user details for security purposes.

3. Data Upload: Users can upload their stock market data for analysis and prediction.

4. Server Verification: The uploaded data is verified by the server, and the relevant information is stored in the database.

5. Prediction Analysis: The system utilizes the Naïve Bayes algorithm to perform prediction analysis based on the uploaded data and user-defined thresholds.

6. Automation for Purchase or Sale: The system calculates the value of shares based on real-time data and notifies users for potential purchase or sale opportunities.

7. Threshold Setting: Users can set their threshold values for shares, and the system generates alert messages based on the equity market movement.

The proposed system aims to enhance the speed and accuracy of stock market prediction and analysis. By leveraging the parallel processing capabilities of Hadoop and the probabilistic modeling of Naïve Bayes algorithm, the

system provides transparent investment opportunities and helps inexperienced investors make informed decisions

IV. UNDERSTANDING OF PROPOSED SCHEME AND MATHEMATICAL MODEL

The proposed scheme in the research paper focuses on utilizing Hadoop and the Naïve Bayes algorithm for stock market prediction and analysis. The scheme consists of two key components: the proposed architecture and the mathematical model.

1. Proposed Scheme:

The proposed architecture incorporates Hadoop and Naïve Bayes algorithm to enable efficient and accurate stock market prediction. It includes the following subtopics:

a. NBE Learning Algorithm: The scheme utilizes the NBE (Naïve Bayes Estimator) learning algorithm. It involves fetching a specified number of previous day's data and calculating the mean values of relevant parameters.

b. Classification: The mean values are used to classify the current stock market values as either above or below the mean. This classification helps determine the probability of the stock value being in a specific category.

c. Bayesian Classifiers: The scheme applies Bayesian classifiers, which utilize Bayes' theorem, to estimate the probability of a particular stock value belonging to a certain class. This calculation considers various factors such as the probability of generating the instance given the class and the probability of occurrence of the class itself.

d. Threshold Calculation: The scheme calculates threshold values based on the maximum and minimum values, the open value of the stock, and the probability of the stock being above or below the mean. These thresholds are used for decision-making in stock purchases or sales.

2. Mathematical Model:

The mathematical model described in the research paper is based on the following subtopics:

a. Components of the Model: The mathematical model is represented by the variables S, E, X, Y, F, DD, NDD, Success, and Failure. These variables define the initial and final states, input and output parameters, functions, deterministic and non-deterministic data, success, and failure conditions of the model.

b. Input for Prediction: The model considers user inputs such as the name of the company for prediction, the number of days for prediction, and the rough amount the user wishes to invest.

c. Deterministic and Non-Deterministic Data: The model distinguishes between deterministic data (such as the number of previous day's requests, company name, and investment amount) and non-deterministic data (such as the open value for each new day).

d. Success and Failure Conditions: The model defines the desired output as successful if the user receives a predicted value within their budget and makes a profitable transaction. On the other hand, if the desired output is not achieved or if the user incurs a loss, it is considered a failure.

The proposed scheme and mathematical model aim to enhance stock market prediction accuracy, automate investment decisions, and provide inexperienced investors with a sound investment strategy.

V. CONCLUSION

After analyzing the research mentioned in the paper, it can be concluded that the proposed system for stock market prediction and analysis, utilizing Hadoop and the Naïve Bayes algorithm, shows promise in enhancing the accuracy and performance of investment decision-making. The system leverages big data processing techniques, data mining, and parallel processing to handle the complexities of analyzing large-scale financial data.

By integrating Hadoop's distributed processing capabilities and the Naïve Bayes algorithm's probabilistic modeling, the system aims to provide transparent investment opportunities and assist inexperienced investors in making sound investment decisions. The use of parallel processing helps overcome delays and makes the system more fault-tolerant.

Although the paper claims an accuracy rate of 70-80% in stock market prediction, further research and empirical testing are necessary to validate these claims and assess the system's real-world performance. Additionally, the system's effectiveness may depend on the quality and relevance of the input data, as well as the adaptability of the Naïve Bayes algorithm to changing market conditions.

Overall, the research presented in the paper highlights the potential of combining big data processing, distributed file systems, and machine learning algorithms for stock market prediction and analysis. It provides a foundation for future studies and advancements in this field, ultimately aiming to support investors in making informed and profitable investment decisions.

REFERENCES

[1] "Data mining with big data" by Xindong Wu et al. (IEEE Transactions on Knowledge and Data Engineering, 2014): This paper explores the challenges and techniques of data mining in the context of big data, providing insights into handling large datasets and extracting meaningful patterns.

[2] "Stock exchange forecasting using Hadoop Map-Reduce technique" by Kushagra Sahu et al. (International Journal of Advancement in Research and Technology, 2013): The authors investigate the application of Hadoop's Map-Reduce technique for stock exchange forecasting, highlighting the potential benefits of distributed processing in analyzing stock market data.

[3] "Data Mining for Big Data: A Review" by Bharti Thakur and Manish Mann (International Journal of Advanced Research in Computer Science and Software Engineering, 2014): This review paper focuses on data mining techniques for big data, providing an overview of various methods and algorithms employed in extracting valuable insights from large datasets.

[4] "Improving Data Transfer Rate and Throughput of HDFS using Efficient Replica Placement" by Neha M Patel et al. (International Journal of Computer Applications, 2014): The authors present a study on enhancing the data transfer rate and throughput of Hadoop Distributed File System (HDFS) through efficient replica placement techniques, improving the overall performance of HDFS.

[5] "Performance Evaluation Of Association Mining In Hadoop Single Node Cluster With Big Data" by A. Asbern et al. (2013 International Conference on Circuit, Power and Computing Technologies): This research paper evaluates the performance of association mining in a Hadoop single node cluster environment with big data, assessing the efficiency and effectiveness of mining frequent patterns.

[6] "Naive Bayes Models for Probability Estimation" by Daniel Lowd and Pedro Domingos (Department of Computer Science and Engineering, University of Washington, 2005): The authors discuss Naive Bayes models for probability estimation, exploring the principles and applications of this probabilistic classifier in various domains, including stock market prediction.

[7] "Stock Price Prediction Using News Articles" by Qicheng Ma (CS224n, Stanford University): This paper focuses on stock price prediction using news articles, investigating the correlation between news sentiment and stock market trends, and exploring the potential of natural language processing techniques in predicting stock prices.

[8] "Mulicluster Hadoop Distributed File System" by Tomasic et al. (IEEE Conference, 2015): The authors present a study on Mulicluster Hadoop Distributed File System, discussing its architecture, features, and benefits in managing and processing large-scale data in a distributed environment.

[9] "Parallel and quantitative sequential pattern mining for large-scale interval-based temporal data" by Guangchen Ran and Hui Zhang (2014 IEEE Conference): This research paper explores parallel and quantitative sequential pattern mining techniques for large-scale interval-based temporal data, addressing the challenges and proposing efficient mining approaches.

[10] "Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data" by J. Josephamenandas and J. Jakkulin Joshi (Global Journal of Advanced Research): The authors investigate data mining techniques with parallel processing for complexity reduction and characterization of big data, focusing on improving the efficiency and effectiveness of data mining algorithms in handling large datasets.