Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 7
July 2023

Optimized Bi-GRU model-basedStock Market Prediction: Bigdata Consideration of Stock and News Data

Shilpa B L[1], Shambhavi B R[2]
[1]Department of Information Science and Engineering, GSSS Institute of Engineering and Technology for Women (GSSSIETW),
Vishweshwariah Technological University (VTU), Karnataka, India
[2]Department of Information Science and Engineering, BMS College of Engineering (BMSCE), Bengaluru, Karanataka, India
Email: shilpabl@gsss.edu.in, blshilpa@gmail.com

Abstract- Sentiment analysis plays a vital role in making informed decisions about business investments in stock markets. It is crucial for identifying an organization's or company's business through stock analysis. Predicting stock prices can be challenging due to their unstable nature, influenced by various factors such as politics, economics, and leadership changes. Historical or textual data alone may not be enough for efficient prediction. Incorporating news sentiment data with stock price data can significantly improve the accuracy of predictions. To this end, we have developed a prediction framework that utilizes both stock price and news sentiment data. The framework initially retrieves technical indicator-dependent features such as Moving Average Convergence Divergence (MACD), Moving Average (MA), and Relative Strength Index (RSI) from stock data. The news data then undergoes specific processes such as pre-processing, feature extraction, and categorization to identify sentiments. In the final categorization stage, the final prediction takes place by the Optimized Bi-GRU model, where the training is carried out by hybrid optimization model that includes Pelican optimization. We conducted a parametric and non-parametric analysis of our proposed POA by altering the parameters.

Keywords*:* PelicanOptimization Algorithm (POA)**,** Stock price prediction**,** Hadoop**,** Big Data**,** Relative Strength Index (RSI)

1. INTRODUCTION (10 PT)

Stock market prices are influenced by numerous factors, including the reaction of investors to financial news and day-to-day events [1]. In recent times, the availability of news has significantly increased, making it difficult for investors to determine stock price trends. Consequently, the development of an automated system capable of predicting future stock prices would be beneficial. Such a system could collect real-time financial news related to the companies of interest and apply a machine learning model to this data, along with historical stock price information, to forecast prices. Researchers have long studied the prediction of stock prices using either historical stock price data or a combination of textual and historical data [2] [3]. Some of the previous works used Twitter sentiments, financial blogs, or news articles as the textual data.

In 2020, Li et al. [4] introduced a stock market forecasting strategy depend on stock prices along with news sentiments. This technique conducted a technical analysis which depicts the price levels via technical indicators and established a new deep learning model in order to increase the knowledge about innovative sentiments as well as technical indicators. Also, a fully linked NN was established for stock prediction. A comparative performance analysis was also made to prove the superiority.

In 2020, Sergio et al. [5] developed a kernel adaptive filtering (KAF) by involving the interdependent system of stock market in the stock return forecasting. This double stage stock return model includes the processes such as sequence modelling and learning the local system's market interdependence from multiple stock markets. In order to increase the predictability, KAF's extant formulations, local models along with local features were gathered from several other stocks.

In 2020, Bouktif et al. [6] introduced an augmented textual feature dependent stock market forecasting method. This method experimentally analyses the stock market

predictability with the usage of expanded format of sentiment analysis. In this work N-grams, sentiment polarity, stock price background as well as subjectivity etc, was utilized to compare the functionality. The adopted method's performance showed precise prediction.

This paper intends to propose a new stock market prediction model that includes both the stock data and news data. Here, both the stock and news data are processed first. The data gathering is done under big data perspective. From the stock data input, features like Improved Simple Moving Average (SMA), Relative Strength Index (RSI), Exponential Moving Average (EMA) [7] [8] [9] are considered. Initially, the news data is subjected to pre-processing that includes keyword extraction and sentiment categorization. Subsequently, the feature like Correlation, Improved Semantic similarity are extracted. Features of both the stock data and News data are fused together to define the final prediction. For this, Improved feature level fusion will take place. The final prediction takes place by the Optimized Bi-GRU model, where the training is carried out by hybrid optimization model that includes Pelican optimization[10] [11].

2. PROPOSED METHOD

Investors aim to forecast market behaviour to make informed decisions while buying or selling stocks, with the ultimate goal of making a profit.

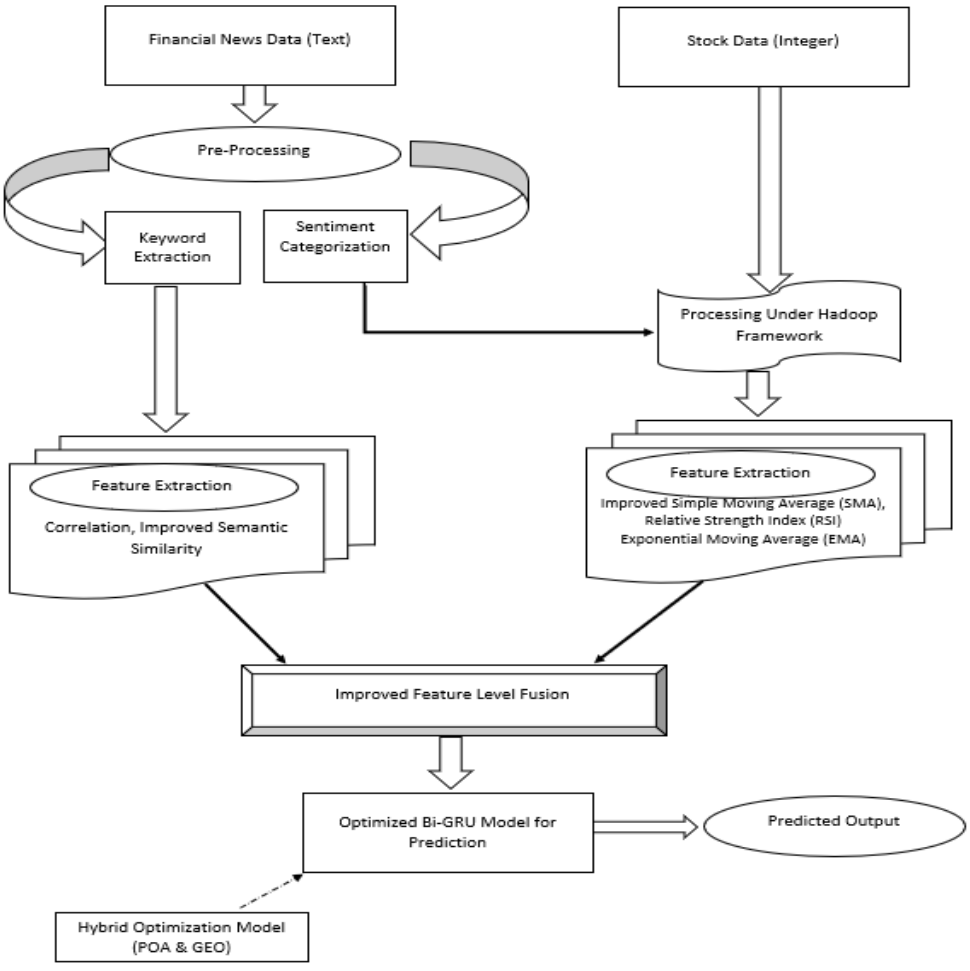

Figure 1: Architecture of Proposed System

However, this is challenging due to the unpredictable nature of market sentiments influenced by various factors like politics, the global economy, and investor expectations. Nevertheless, investor sentiment is more visible due to herd-like behaviour, overreaction, and limited institutional engagement. To determine whether emotions or logical decision-making are driving the stock market, investors use sentiment research. They analyse unstructured data like financial news articles or headlines to extract the necessary information. Many experts

suggest using text mining and machine learning techniques to examine the text and extract information relevant to forecasting. Several studies in machine learning, deep learning, and natural language processing (NLP) have been conducted in this area, including extracting and analysing opinions from media headlines about the stock market and predicting a company's specific stock performance based on its sentiment. The Figure 1 above shows the proposed system architecture.

2.1 EXTRACTION OF FEATURES FROM FINANCIAL NEWS DATA

a. PRE-PROCESSING

Pre-processing refers to the techniques and methods used to prepare data for analysis or machine learning tasks. It involves cleaning and transforming raw data into a format that is suitable for analysis, which can help to improve the accuracy and efficiency of data analysis and machine learning models. Pre-processing can include a variety of tasks such as data cleaning, data transformation, data integration, data reduction, and data normalization.

Improved Stemming: It is a technique to eliminate suffix of word and it aids in minimizing the needed number of computations [14]. Stemming is deployed in diverse appliances like information retrieving systems. Moreover, it is exploited in domain analysis for determining the domain vocabulary. In this work, improved stemming process is performed. However, there is no checking of dictionaries. This has an impact on all words that goes to stemming process. Here, we have added dictionary checking process in first step. After that based upon concurrence analysis, the stemming process is done.

In Eq. (1), the expected mutual information between two terms, a and b, is modelled using various parameters. The count of times that co-occur in a fixed-size text window is denoted by n_{ab}, while the count of occurrences of a and b in the corpus is denoted by $EM(a,b)$ and k refers to constant based upon corpus and window size. The computation of this is shown in Eq. (2).

$$EM(a,b) = \max\left(\frac{n_{ab} - EM(a,b)}{K * n_a + n_b}, 0\right) \quad (1)$$

$$K = \frac{\sum n_{ab}}{\sum n_a n_b} \quad (2)$$

Tokenization: This process converts the text into tokens prior to transferring to vectors [15]. Tokenization is a straightforward process that involves breaking down raw text into smaller chunks known as tokens. During this process, sentences and words are identified and labelled as tokens. It is also easy to filter out any unnecessary tokens, which can help in preparing the text for further analysis or machine learning tasks. Subsequently, the tokens help to know the context for NLP and aids in understanding the meaning of text via examining the sequence of words.

Keyword Extraction:

Keyword extraction, also referred to as keyword detection or keyword analysis, is a technique used in text analysis that automatically identifies the most frequently used and significant words and phrases from a given text [16]. This technique aids in summarizing the content of the text and identifying the primary topics discussed. Use stop words to filter common word such as they, we, have, etc. We don't need that because it's not really important for a keyword. Vectorize and transform the words using TF-IDF function. Get keywords using TF-IDF function and save it.

2.2 EXTRACTION OF FEATURES FROM STOCK DATA VIA HADOOP FRAMEWORK

Hadoop is a distributed computing framework that enables the processing of large datasets across multiple machines in a cluster. It is designed to handle big data and can perform computations at a scale that would not be possible with traditional computing architectures. One important feature of Hadoop is its ability to perform feature extraction, which involves identifying and selecting the most relevant features from a dataset. Feature extraction is a critical step in machine learning and data analysis, as it helps to reduce the dimensionality of the data and improve the accuracy of models. Hadoop offers a range of tools and libraries that can be used for feature

extraction, including Apache Mahout, which provides algorithms for clustering, classification, and collaborative filtering. Other tools such as Apache Spark and Apache Flink can also be used for feature extraction and machine learning tasks. In summary, Hadoop is a powerful framework for big data processing that can be used in conjunction with various tools and techniques for feature extraction and machine learning. It provides a scalable and efficient platform for data analysis that is increasingly being used in industry and academia.

A. Hadoop Frame work:

Hadoop provides a definite file system, known as HDFS, for organizing processed data, which might also be scalable, reliable and distributed. When considering "Map and Reduce", the fundamental step is data fragmentation and the fragmented data is again provided to a distributed network for processing purposes. This processed data is incorporated into one. A Hadoop scheme considers node failure, which is automatically handled and this makes Hadoop quite versatile with a flexible platform deployed for a variety of demanding appliances.

From the hadoop based big data, the features namely, ATR, Bollinger bands, ADI and EOM are derived.

B. Feature Extraction:

a. Improved Simple Moving Average (SMA)

SMA is a commonly used technical analysis indicator in finance that calculates the average price of a security over a specified period of time. It is often used to identify trends and potential trading opportunities.

$$SMA = \frac{A1+A2+\ldots+An}{n} (3)$$

where:

An=the price of an asset at period *n,*

n=the number of total periods

The SMA method aids traders in forecasting the direction of price changes. When the SMA increases, it indicates an upward trend, whereas a decreasing SMA suggests a downward trend. Additionally, a short-term SMA crossing above a long-term SMA indicates an expected uptrend, whereas a long-term SMA surpassing a short-term SMA suggests a possible downtrend.

b. Relative Strength Index (RSI)

The relative strength index is a momentum-based indicator that measures a security's strength during upward and downward price movements. By analysing the relationship between this measure and price action, traders can gain insights into the potential performance of a security. When combined with other technical indicators, the RSI can assist traders in making more informed decisions when trading.

Calculating RSI

The RSI uses a two-part calculation that starts with the following formula given below:

$$RSIstep\ one = 100 - \left[\frac{100}{1+\frac{Average\ gain}{Average\ loss}}\right] (4)$$

The calculation for average gain or loss involves determining the average percentage gain or loss over a specific look-back period. When calculating this formula, a positive value is assigned to the average loss. Any periods with price losses are treated as zero in the calculation of average gain, while periods with price increases are counted as zero in the calculation of average loss.

c. Exponential Moving Average (EMA)

The Exponential Moving Average (EMA) is a technical tool utilized in trading to track the price changes of an asset or security within a specified period. It is a weighted moving average (WMA) that places greater importance on recent price data, making it more responsive to changes. Similar to the Simple Moving Average (SMA), the EMA is employed to identify trends in price over time. A Moving Average Ribbon can be used to monitor several EMAs simultaneously.

When it comes to calculating the EMA of a stock:

$$EMA = \text{Price}(t)\times k + EMA(y)\times(1-k) (5)$$

where:

t=today, y=yesterday, N=number of days in EMA, $k=2\div(N+1)$

C. Improved Feature Level Fusion:

Information fusion is an advanced technology that gathers various data from multiple sensors on a single object and eliminates noise or redundant information from mutual data. The fusion process typically involves three levels: data level fusion, feature level fusion, and decision level fusion. Feature level fusion is widely used in image recognition and fault diagnosis due to its simplicity. This approach

involves extracting multiple types of features from the original data and combining them using fusion techniques. There are three fusion forms: series fusion, parallel fusion, and complex vector fusion, which have been utilized in various research fields. Feature fusion is a valuable method for combining different features into a unified feature for the same recognition problem. The advantage of feature fusion is that it retains useful information about the original features while also reducing redundancy to some extent.

In this research work, we introduce a stock market predictive framework that utilizes feature fusion. The framework consists of three modules: an auto-regression module that extracts feature A from historical closing price data, a multi-variable regression module that extracts feature B from related technical variables, and a feature fusion module that combines features A and B using specific fusion methods to create a unified feature. This unified feature serves as the input for prediction tools like ANNs or SVR, with the predicted future closing price being the output.

2.3 OPTIMIZED BI-GRU FOR PREDICTING STOCK MARKET

Optimized BI-GRU (Bidirectional Gated Recurrent Unit) is a state-of-the-art deep learning model used for predicting stock market trends. This model is designed to analyse the historical data of a stock and identify patterns that can help predict future price movements. By utilizing a bidirectional approach, the model can consider both past and future data points when making predictions. In addition, optimization techniques such as dropout and batch normalization are employed to enhance the model's accuracy and prevent overfitting. The resulting model has shown promising results in predicting stock prices and is increasingly being used by traders and investors as a valuable tool in their decision-making process.

1. Hybrid Optimization Model

The Hybrid Optimization Model (HOM) is a mathematical modelling technique that combines different optimization algorithms to solve complex problems. It is an iterative process that uses multiple algorithms to find the best solution for a given problem. The idea behind HOM is to combine the strengths of different optimization algorithms, such as genetic algorithms, particle swarm optimization, simulated annealing, and others, while minimizing their weaknesses. This approach can lead to better solutions than using a single algorithm alone.

2. Pelican Optimization Algorithm (POA)

The Pelican Optimization Algorithm (POA) is a swarm-based optimization algorithm that is inspired by the hunting behavior of pelicans. Pelicans are known for their ability to work together in groups to catch prey, and the POA algorithm mimics this behavior to find optimal solutions for complex problems. During hunting, pelicans dive from a height and spread their wings on the water's surface to force the fish to move towards shallow waters. They then use their beak to catch the fish and remove excess water before swallowing it. This process involves intelligent decision-making and coordination among the pelicans. The POA algorithm is designed based on this behavior and strategy of pelicans. It uses a swarm of agents to explore the search space and find optimal solutions. The agents move towards promising areas of the search space and communicate with each other to share information and coordinate their movements. This allows the algorithm to efficiently search for the best possible solutions. In summary, the POA algorithm is inspired by the intelligent hunting behavior of pelicans and uses a swarm of agents to find optimal solutions for complex problems. It is a promising approach for optimization and has been successfully applied in various fields, including engineering and finance.

Mathematical Model of the Proposed POA

The Pelican Optimization Algorithm (POA) is a type of population-based algorithm that incorporates pelicans as members of the population. In population-based algorithms, each population member represents a potential solution to the optimization problem. They propose values for the problem variables based on their position in the search space. At the start

of the algorithm, population members are randomly initialized within the problem's bounds using Equation given below.

$$x_{i,j} = l_j + rand \cdot (u_j - l_j),\ i=1,2,\ldots,N,\ j=1,2,\ldots,m,$$

where $x_{i,j}$ is the value of the *j*th variable specified by the *i*th candidate solution, N is the number of population members, m is the number of problem variables, rand is a random number in interval [0, 1], l_j is the *j*th lower bound, and u_j is the *j*th upper bound of problem variables.

The POA algorithm mimics the behavior and tactics of pelicans during their hunt for prey to update candidate solutions. This hunting strategy is emulated through two stages:

- **Moving towards prey (exploration phase)**

During the first phase, the pelicans locate the prey and navigate towards it. The POA algorithm emulates this strategy by scanning the search space and exploring different areas to locate the optimal solution. It is important to note that in the POA algorithm, the prey's location is randomly generated within the search space to increase its exploration power in the problem-solving space. The POA algorithm employs effective updating, meaning that the new position of a pelican is only accepted if it improves the objective function's value. This helps prevent the algorithm from moving towards non-optimal areas, ensuring that it stays on track towards the optimal solution.

- **Winging on the water surface (exploitation phase)**

During the second phase, the pelicans spread their wings on the water's surface to move the fish upwards and collect them in their throat pouch. This behaviour enables them to catch more fish in the attacked area. The POA algorithm emulates this behaviour to converge towards better solutions in the search space, increasing its local search power and exploitation ability. Mathematically, the algorithm examines the points in the pelican's neighbourhood to converge towards a better solution. This process involves exploring the search space in the vicinity of the current solution to identify the best possible solution. Overall, the second phase of the POA algorithm helps it to converge towards the optimal solution efficiently.

The pseudo code of the adopted POA method is represented in Algorithm 1.
Start POA.
1. Input the optimization problem information.
2. Determine the size of the POA population (N) and the number of iterations (T).
3. Initialize the positions of the pelicans randomly and calculate the objective function for each member.
for t = 1 to T
4. Generate the position of the prey randomly.
for i = 1 to N
5. Phase 1: Moving towards prey (exploration phase)
For j = 1 to m
6. Calculate the new value of the j-th dimension.
End.
7. Update the i-th population member.
8. Phase 2: Winging on the water surface (exploitation phase)
For j = 1 to m
9. Calculate the new value of the j-th dimension.
End.
10. Update the i-th population member
End.
11. Update the best candidate solution found so far.
End.
Output the best candidate solution found by POA.
End POA.

3. RESULTS AND DISCUSSION

The Proposed method based on Stock Market Prediction was implemented in Python and the obtained results were analysed. To evaluate the proposed method, we had utilized five different types of datasets and it was collected from [17]. Additionally, the analysis was carried out with respect to MAE, MSE and RMSE by varying the learning percentage to 60, 70, 80 and 90, respectively. The suggested strategy will be evaluated by varying the parameter 'l'.

The stock dataset includes two companies such as Reliance Communications and Relaxo Footwear. In addition, each company consists of three datasets (a) in daily option, set start day 1-1-2019 and end day 1-12-2020, (b) in monthly option, set start jan2000 and end dec2020, and (c) in yearly option, set year 2000.

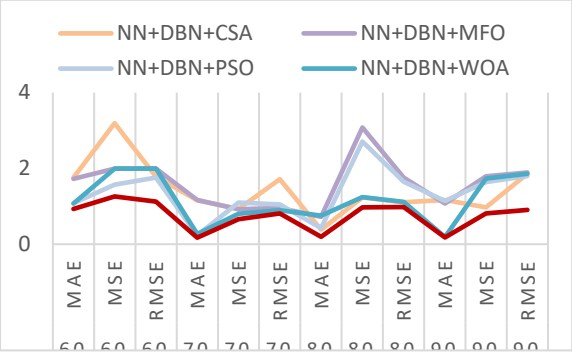

Figure 2: Performance Analysis on error measure with respect to MAE, MSE and RMSE

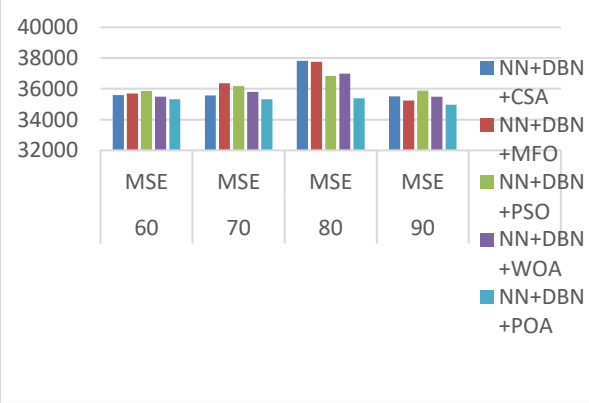

Figure 3: Performance Analysis on error measure with respect to MSE on data set 5

The performance of proposed model regarding error metrics like MAE, MSE and RMSE is explored in this division. Here, analysis was done on six datasets and the outcomes are displayed in figure 2 and figure 3. The adopted method obtained very low MAE, MSE and MSLE in almost all the learning percentage. Also, it was analyzed by varying the parameter to l=-0.5, l=0.5 and l=0.8. According to dataset1, while varying the parameter l to -0.5, the proposed method attains the MAE of 0.92 in the 90% of learning rate. Next, in the variation of 0.5 and 0.8, the proposed method has maintained the MAE as 0.88 and 0.90. For the MSE analysis, at the 60% of learning percentage, the minimal MSE of the proposed method under l variation of 0.5 is 1.14. It is observed that 1.21 of MSE are finally maintained by the proposed method under the variation of l=0.8. Also, it exhibits the lowest RMSE error value, ranging from 1.03 to 1.35, under the variation 0.8. Analyzing the dataset2, in the 60% of learning rate, the l under the variation -0.5, 0.5 and 0.8, get the lower RMSE as 187, 180 and 184. Next, the lowest MSE is attained in the 80% of learning percent as 35000 (under the variation 0.5).

Moreover, the proposed method holds the minimum error value as (~) 27, 38, 42 and 41 at the learning percentage 60, 70, 80 and 90, respectively for dataset5 (under the variation l=0.8). Similarly, under the variation -0.5, the proposed method yields the lower MAE error rate, ranges from 26 to 43. Based on dataset6, the suggested method obtains the RMSE is 396, 398 and 400, under the variation of l is -0.5, 0.5 and 0.8. This has proved that the proposed work obtained with lesser error even under different variations and it is suitable for stock market prediction.

4. CONCLUSION

Stocks are traded, exchanged, and circulated in the stock market, with investors focusing heavily on the fluctuating pattern of stock prices. These fluctuations are often nonlinear, making it critical for economists to forecast them accurately. To address this need, we have developed a prediction framework that uses sentiment analysis to analyze both stock price and news sentiment data. Initially, technical indicator-dependent features including Simple Moving Average (SMA), Exponential Moving Average (EMA) as well as Relative Strength Index (RSI) was retrieved from those stock data. To identify sentiments from those news data, we conducted specific processes such as pre-processing, feature extraction as well as categorization. In the final categorization stage, to provide a precise sentiment prediction the utilized Deep Belief Network (DBN) weights was tuned by Pelican Optimization algorithm (POA). A parametric and non-parametric analysis was conducted in this work for POA via altering the parameters and the outcomes indicated that proposed POA can offer superior performance compared to other algorithms.

REFERENCES

[1] Lu, W., Li, J., Wang, J. and Qin, L., 2021. A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 33(10), pp.4741-4753.
[2] Dai, Z., Zhou, H., Wen, F. and He, S., 2020. Efficient predictability of stock return volatility: The role of stock market implied volatility. The North American Journal of Economics and Finance, 52, p.101174.
[3] Kumari, J., 2019. Investor sentiment and stock market liquidity: Evidence from an emerging economy. Journal of Behavioral and Experimental Finance, 23, pp.166-180.
[4] Li, X., Wu, P. and Wang, W., 2020. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), p.102212.
[5] Garcia-Vega, S., Zeng, X.J. and Keane, J., 2020. Stock returns prediction using kernel adaptive filtering within a stock market interdependence approach. Expert Systems with Applications, 160, p.113668.
[6] Bouktif, S., Fiaz, A. and Awad, M., 2020. Augmented textual features-based stock market prediction. IEEE Access, 8, pp.40269-40282.
[7] Guo, K., Sun, Y. and Qian, X., 2017. Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. Physica A: Statistical Mechanics and its Applications, 469, pp.390-396.
[8] Duan, J., Luo, B. and Zeng, J., 2020. Semi-supervised learning with generative model for sentiment classification of stock messages. Expert Systems with Applications, 158, p.113540.
[9] Vinolin, V. and Vinusha, S., 2018. Enhancement in biodiesel blend with the aid of neural network and SAPSO. Journal of Computational Mechanics, Power System and Control, 1(1), pp.11-17.
[10] Pavel Trojovský and Mohammad Dehghani, "Pelican Optimization Algorithm: A Novel Nature-Inspired Algorithm for Engineering Applications", Sensors, 2022.
[11] Abdolkarim Mohammadi-Balani, Mahmoud Dehghan Nayeri, Adel Azar, Mohammadreza Taghizadeh-Yazdi, "Golden eagle optimizer: A nature-inspired metaheuristic algorithm", Computers & Industrial Engineering, vol. 152, 2021.
[12] Derakhshan, A. and Beigy, H., 2019. Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, pp.569-578.
[13] Chen, Y., Lin, W. and Wang, J.Z., 2019. A dual-attention-based stock price trend prediction model with dual features. IEEE Access, 7, pp.148047-148058.
[14] Chen, Y., Lin, W. and Wang, J.Z., 2019. A dual-attention-based stock price trend prediction model with dual features. IEEE Access, 7, pp.148047-148058.
[15] Wu, S. and Wang, S., 2011. Information-theoretic outlier detection for large-scale categorical data. IEEE transactions on knowledge and data engineering, 25(3), pp.589-602.
[16] Mohan, Y., Chee, S.S., Xin, D.K.P. and Foong, L.P., 2016, December. Artificial neural network for classification of depressive and normal in EEG. In 2016 IEEE EMBS conference on biomedical engineering and sciences (IECBES) (pp. 286-290). IEEE.
[17] https://www.moneycontrol.com/stocks/histstock.php