Journal of Theoretical and Applied Information Technology

15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ELITE ARTIFICIAL BEE COLONY OPTIMIZATION-BASED SYNERGY RANDOM FOREST (EABC-SRF) FOR ADDRESSING AMBIGUITY IN STOCK TWEETS

[1]G.PRIYADARSHINI, [2] Dr.D,KARTHIKA

[1]Research Scholar, P.K.R. Arts College for Women,
[2] Associate Professor and Head, Department of Computer Science,
VET Institute of Arts and Science (Co-education) College, Erode, Tamil Nadu, India

E-mail: [1] savidpri@gmail.com, [2]karthikad@vetiac.ac.in

ABSTRACT

The stock market is a complex and dynamic financial ecosystem where investors buy and sell securities. Various factors influence it, including economic indicators, geopolitical events, and social sentiment. Twitter has become a significant source of real-time information for traders and investors. Stock tweets are short messages posted on Twitter that discuss stocks, providing insights, opinions, and predictions. Analyzing these tweets can help gauge market sentiment and anticipate price movements. The classification of stock tweets involves categorizing them as positive, negative, or neutral based on sentiment. This sentiment analysis aids in understanding investor sentiment and predicting market trends. The Elite Artificial Bee Colony Optimization-Based Synergy Random Forest (EABC-SRF) is an innovative algorithm to enhance sentiment analysis. It combines the power of Artificial Bee Colony Optimization (ABC) and Synergy Random Forest (SRF) to optimize feature selection and sentiment classification. EABC-SRF uses elite artificial bees to select the most relevant features for sentiment analysis. It then integrates these features into the SRF framework to classify tweets effectively, reducing ambiguity and noise. The "Stock Tweets for Sentiment Analysis and Prediction" dataset is the foundation for training and testing EABC-SRF. It contains a vast collection of stock-related tweets for model development and evaluation. Results from experiments with EABC-SRF demonstrate its superior performance in sentiment analysis compared to traditional methods. It disentangles the ambiguity in stock tweets, providing valuable insights for investors and traders in predicting market sentiment and trends.

Keywords: *Stock Market, Twitter, Classification, Random Forest, Artificial Bee Colony, Optimization*

1. INTRODUCTION

Sentiment analysis, often called opinion mining, is a dynamic field in natural language processing that involves assessing and deciphering the emotional tone of textual content [1]. It empowers businesses, researchers, and individuals to gauge public sentiment toward a particular topic, product, or service, whether positive, negative, or neutral. Using machine learning algorithms and linguistic analysis, sentiment analysis extracts valuable insights from social media posts, customer reviews, and news articles. These insights aid decision-making, brand management, and customer satisfaction [2], [3].

Stock Tweets, known for their brevity, provide a unique lens into stock trading and investment. These concise messages, typically confined to 280 characters, are prevalent on platforms like Twitter and offer real-time insights, opinions, and reactions from traders and investors. One of the critical advantages of Stock Tweets is their ability to swiftly disseminate information about market events [3]. This immediacy makes them a valuable resource for market participants looking to stay up-to-date with breaking news, earnings reports, and notable stock movements. Moreover, the concise format encourages users to distil complex financial ideas into easily digestible snippets, making them accessible to a broader audience. The brevity of Stock Tweets can also pose challenges. It often leads to oversimplification, where critical nuances may be overlooked [4]. Therefore, while they offer a valuable glimpse into market sentiment, investors must

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

conduct comprehensive research and due diligence before making trading decisions based solely on these tweets. Stock Tweets are a rapid, dynamic, and influential component of the modern financial landscape, offering opportunities and pitfalls to those who engage with them [5].

Twitter, a microblogging platform, has become a treasure trove of real-time data for sentiment analysis. With millions of tweets posted daily on various topics, it is a valuable resource for understanding public sentiment [6], [7]. Sentiment analysis on Twitter involves using natural language processing (NLP) techniques to classify tweets as positive, negative, or neutral based on the emotions expressed within the text. This analysis can provide several insights:

- **Brand Monitoring:** Companies can track mentions of their brands and products on Twitter to gauge customer sentiment. This helps in understanding public perception and addressing issues promptly.
- **Political Analysis:** Sentiment analysis on political tweets during elections or key events can provide valuable information about voter sentiment and potential outcomes.
- **Crisis Management:** During natural disasters or public health emergencies, sentiment analysis can help authorities monitor public sentiment to take timely actions and respond to concerns.
- **Customer Feedback:** Businesses can gather feedback from customer tweets, allowing them to improve their products and services.
- **Market Research:** Twitter sentiment analysis can inform market research by identifying trends, preferences, and emerging industry topics.

Analyzing sentiment on Twitter comes with challenges like sarcasm, context, and brevity. Researchers and analysts continually refine NLP models to tackle these challenges and extract meaningful insights from the vast sea of tweets, making Twitter a valuable source for sentiment analysis in today's digital landscape.

Bio-inspired optimization, a field rooted in mimicking natural processes and principles, draws inspiration from the complexity and efficiency of biological systems to solve complex problems. Instead of focusing on specific algorithms, bio-inspired optimization harnesses the underlying concepts found in nature to tackle various computational challenges [8]. One fundamental aspect of bio-inspired optimization is its ability to adapt and evolve. Nature demonstrates this through evolution, where species adapt to changing environments over time. Similarly, bio-inspired optimization methods are designed to adapt their strategies or parameters to find optimal solutions in dynamic or uncertain problem domains [9] .

Decentralization is another critical principle in biological systems because complex behaviours often emerge from the interactions of individual components, such as cells or ants in a colony[10]. This decentralized approach is mirrored in bio-inspired optimization, where a population of solutions collaboratively explores the solution space, allowing for a distributed search that can efficiently navigate complex landscapes. Bio-inspired optimization can explore diverse solution spaces akin to genetic diversity in species. This diversity allows it to avoid getting trapped in local optima and instead seek global optima, ensuring better overall solutions to complex problems. Significant advantages of bio-inspired optimization in sentiment classification on Twitter are [11]:

- **Adaptability:** Twitter data is dynamic and constantly evolving, with shifts in language and sentiment trends. Bio-inspired optimization can adapt to these changes quickly, ensuring the sentiment classification model remains effective.
- **Distributed Processing:** Twitter generates vast amounts of data from diverse sources and languages. Bio-inspired optimization's decentralized approach can efficiently handle this distributed data, making it well-suited for sentiment analysis on Twitter.
- **Global Optimization:** Sentiment analysis requires a nuanced understanding of evolving language expressions. Bio-inspired optimization's ability to explore diverse solution spaces can help find the best model configurations for accurate sentiment classification.

Bio-inspired optimization's [12], [13], [22]–[26], [14]–[21] adaptability, decentralized processing, and capacity for global optimization make it a promising approach for sentiment classification on Twitter, enabling the development of robust models capable of tackling the platform's dynamic and multilingual nature.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1.1. Problem Statement

In the domain of sentiment classification for stock tweets, a prominent and persistent challenge emerges from the inherent vagueness and subtleties present in the language used, including the occasional presence of sarcasm, which collectively perplex sentiment analysis algorithms and hinder their capacity to discern sentiment polarity. The nature of stock tweets, characterized by their conciseness and informality, frequently results in adopting succinct and sometimes enigmatic language, thus obscuring the intended sentiment. Moreover, the frequent deployment of sarcasm in stock-related tweets adds another intricate layer of complexity, as it often conveys sentiments instead of the words' literal meaning. These multifaceted intricacies underscore the necessity for pioneering approaches that can adeptly decode the intricate linguistic subtleties and resolve the intricate challenge of achieving precise sentiment classification within the ever-evolving landscape of financial markets.

1.2. Motivation

The impetus for this research in sentiment classification within stock tweets is grounded in the rapidly evolving landscape of social media's impact on financial markets. In today's digitally interconnected world, platforms like Twitter serve as dynamic market sentiment indicators, wielding the power to swiftly influence investment decisions and market dynamics. However, the pervasive challenge of linguistic ambiguity and the subtleties of sarcasm inherent in stock tweets obstruct precise sentiment analysis. This research is driven by the quest to devise innovative approaches for disentangling the intricate language of stock tweets, ultimately advancing the accuracy of sentiment classification models. The significance of practical sentiment analysis transcends investor empowerment to encompass a broader understanding of the intricate interplay between social media sentiment and stock price movements, promising benefits not only for financial professionals but also for a broader spectrum of market participants seeking informed decision-making in an era where social media profoundly shapes market sentiment.

1.3. Objectives

The central research objective is to enhance sentiment classification accuracy within stock tweets by effectively tackling the challenges posed by linguistic ambiguity and sarcasm. This study aims to quantify and analyze the extent of linguistic ambiguity, identify its prevalent patterns, and develop data preprocessing strategies to disambiguate and enrich stock tweets' context. Furthermore, the research will focus on developing and implementing advanced machine learning and natural language processing models tailored to decode the intricacies of language in stock tweets, particularly addressing sarcasm detection. Through rigorous evaluation, the study aims to demonstrate the efficacy of these models in improving sentiment classification within stock tweets, ultimately providing financial professionals and investors with more reliable tools for decision-making while contributing to a deeper understanding of how linguistic subtleties impact market sentiment in the financial domain.

2. LITERATURE REVIEW

The "Digital Haves and Have-Mores" [27] presents an innovative approach that utilizes social media data to forecast the stock market's dramatic fluctuations during the pandemic. This research goes beyond traditional market analysis by focusing on the digital 'haves'—individuals and entities with significant online presence and influence—and the 'have-mores' with greater access to data and resources. The "Twitter-Aided Decision Making" [28] provides an insightful overview of the latest advancements in utilizing Twitter data for decision-making across various domains. This comprehensive analysis delves into the innovative ways Twitter has been harnessed as a valuable source of real-time information and sentiment analysis.

The "Social Media with Stock Price Movements" [29] study comparing traditional news and social media with stock price movements aims to unravel whether news or price changes occur first. While the Efficient Market Hypothesis posits that news and price adjustments happen simultaneously in highly efficient markets, empirical evidence suggests that there can be lags in information processing. The "Social Informedness and Investor Sentiment" [30] investigates the role of social media and investor sentiment during the GameStop short squeeze phenomenon. This research delves into "social informedness," which refers to how information and sentiment are disseminated and absorbed within online communities, particularly on platforms like Reddit and Twitter.

The "Social media, political uncertainty, and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

stock markets" [31] is a dynamic and influential relationship. Social media platforms serve as lightning-fast conduits for disseminating information, shaping market sentiment, and amplifying reactions to political events and policy changes. Investors increasingly monitor these platforms for real-time insights into political developments that can impact market behaviour. The "Fundamental and Technical Analysis" [32]critically examines the effectiveness and implications of two prominent approaches in stock market forecasting: fundamental analysis and technical analysis. Fundamental analysis involves evaluating a company's financial health, including earnings, assets, and market position, to predict stock prices. Technical analysis relies on historical price and volume data, looking for patterns and trends to make predictions.

The "Information Diffusion over Social Network" [33] comprehensively analyses the mechanisms governing information propagation through social media platforms and its relevance to stock market dynamics. It investigates how information spreads across social networks, shaping investor sentiment and influencing stock market trends. The survey outlines potential developments, discussing emerging trends, technologies, and research avenues that could enhance our understanding of information diffusion's impact on financial markets. The "Visuals and Attention" [34] investigates the impact of visual content, such as images and infographics, on user engagement with earnings-related news shared on Twitter. It explores whether the inclusion of visuals enhances user attention and interaction with such financial information by analyzing metrics like likes, retweets, and comments on tweets containing earnings news with and without visual elements.

The "Twitter and the Voluntary Disclosure Effect" [35] delves into the intriguing relationship between companies' voluntary disclosures and the ensuing discussions and reactions on the Twitter platform. It aims to scrutinize how voluntary disclosure, where companies share information about their operations, financial health, or strategic plans, influences investor sentiment and reactions within the Twitterverse. The "Stock Returns and Investor Sentiment" [36] explores the intricate relationship between stock market returns and investor sentiment, focusing on textual analysis and social media platforms. This study employs advanced text analytics techniques to dissect the sentiment and content of messages and discussions shared on social media channels like Twitter, Reddit, or financial forums.

The "Social Media Sentiment Stock Prediction" [37] framework leverages social media sentiment analysis to enhance stock market prediction, particularly emphasizing the effectiveness of LSTM (Long Short-Term Memory) deep learning models in capturing sentiment dynamics. This approach capitalizes on real-time sentiment insights by systematically collecting and preprocessing social media data related to financial markets. "Twitter Opinion Stock Prediction" [38] framework employs customer opinions from Twitter data to drive stock predictions, utilizing the optimized Strawberry-based Bi-directional Recurrent Neural Model (SBRNM). Its operational mechanism encompasses data collection, preprocessing, and advanced deep learning analysis. This approach systematically gathers Twitter data relevant to financial markets and specific stocks, emphasizing customer sentiments. The SBRNM model's advantage lies in its capacity to capture both past and future context, enhancing predictive accuracy.

3. PROPOSED WORK

3.1. Synergy Random Forest

Synergy Random Forest is a machine learning algorithm that combines the principles of two popular techniques: Random Forests and Synergistic Ensemble Learning. The breakdown of the components of this algorithm are:

- **Random Forests:** Random Forest is a widely-used ensemble learning technique in machine learning. It builds multiple decision trees during training and combines their predictions to make more accurate and robust predictions. Random Forests introduce randomness in two main ways: bootstrapping (sampling with replacement) from the training data and considering only a random subset of features at each node when growing each tree. This randomness helps reduce overfitting and improves the generalization ability of the model.
- **Synergistic Ensemble Learning:** Synergistic ensemble learning involves combining multiple models or algorithms to leverage their complementary strengths, aiming to achieve better performance than individual models. In synergy, the combined effect of the ensemble is greater than the sum of its parts. It often involves selecting models that excel in different areas or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

have different biases and combining them effectively.

3.1.1. Data Collection

Consider a dataset D containing n pairs of feature vectors and corresponding target values: $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$. Here, each X_i represents the feature vector for the i-th data point, and y_i is the corresponding target value. This acts as a foundation for the Random Forest model.

3.1.2. Data Preprocessing

For each feature x_{ij} in a given data point (x_i), if it happens to be missing (denoted as NaN), it can be replaced with an imputed value. One common approach is to use the mean (μ_j) of feature x_j across all available data points. Eq.(1) expresses the same.

$$x_{ij} = \mu_j, \text{ if } x_{ij} \text{ is missing} \quad (1)$$

To identify outliers in a specific feature x_j, calculate its mean (μ_j) and standard deviation (σ_j) across all data points. Then, evaluate the Z-score (Z_{ij}) for each observation x_{ij} using Eq.(2).

$$Z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j} \quad (2)$$

In Eq.(2), if the absolute value of Z_{ij} exceeds a predetermined threshold, consider x_{ij} as an outlier. For proper modelling, bringing all features to a similar scale is essential. Eq.(3) expresses Min-Max scaling, which is applied to scale the feature.

$$S_{ij} = \frac{\left(x_{ij} - min(x_j)\right)}{\left(max(x_j) - min(x_j)\right)} \quad (3)$$

Eq.(3) can be standardized by transforming x_{ij} to a standardized value Z_{ij} and it is expressed in Eq.(4).

$$Z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j} \quad (4)$$

3.1.3. Bootstrapping

This phase prepares subsets for constructing individual decision trees within the Random Forest. Starting with the original dataset $D = \{(x_i, y_i)\}$, where i ranges from 1 to n, and for each subset $b = 1, 2, \ldots, B$.

A new subset D_a is generated by randomly selecting n data points (X_i, y_i) with replacement from the original dataset D and the Eq.(5) expresses the same.

$$D_a = \{(X_{i1}, y_{i1}), (X_{i2}, y_{i2}), \ldots, (X_{in}, y_{in})\} \quad (5)$$

A decision tree T_b is construct using the subset D_a. Each tree will learn different perspective due to the random subsets, contributing to the ensemble's diversity.

3.1.4. Reliable Decision Trees Building

This step involves constructing individual decision trees using the bootstrapped subsets and introducing randomness through feature selection during node splits, for each subset $b = 1, 2, \ldots, B$, where B is the total number of decision trees in the Random Forest. Each node in the decision tree considers a random subset of characteristics for branching. This introduces diversity and reduces overfitting. Let's denote the total number of features as p, and the number of features considered at each split as m. Randomly select m features from the available p features.

Given a node in the tree, this research evaluates candidate splits for each feature. The goal is to find the best split that optimally separates the data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

points based on a specific criterion. Eq.(6) is applied for classification tasks: the Gini impurity.

$$Gini(t) = 1 - \sum (pi, t)^2 \quad (6)$$

where pi, t is the proportion of samples of class i at node t.

Eq.(7) is applied for regression tasks, which is the mean squared error (MSE) and acts as a standard criterion.

$$MSE(t) = \sum (y_i - \hat{y}_i)^2 / n_i \quad (7)$$

where y_i is the target value of the i-th sample, $\hat{y}_i$ is the predicted value, and n_i is the number of samples in the node.

Recursive node partitioning continues until some stopping requirement is met, such as the required number of samples per leaf or the maximum depth of the tree. This process creates a hierarchical structure of nodes and leaves, allowing the tree to capture complex relationships in the data. Further data points are handled by branching from the decision tree's root node to a leaf node, again based on the feature values. The target value associated with the majority class (classification) or the average value (regression) of the samples in the leaf node becomes the predicted value.

3.1.5. Voting

After constructing individual decision trees in the Random Forest, this step involves combining their predictions to make a final prediction. Classification and regression stages must be carried out for each new data point.

If the Random Forest is used for classification, each decision tree votes for its predicted class. The class with the majority of votes among all the trees is considered the final prediction for the Random Forest. Eq.(8) mathematically expresses the same: if it has B decision trees, the predicted class for a new data point is the mode (most common value) of the class predictions across all trees.

$$Final\,Prediction = Mode(Class\,P \quad (8)$$

If the Random Forest is used for regression, the predicted values from all individual trees are averaged to obtain the final prediction. For B number of decision trees, the final predicted value for a new data point is the mean of the predicted values from all trees, and it is mathematically expressed as Eq.(9).

$$Final\,Prediction = \left(\frac{1}{B}\right) \times \sum Predic \quad (9)$$

3.1.6. Ensemble Output

The output of the Random Forest algorithm is now ready. It provides predictions for new data points based on the combined decisions of the ensemble of decision trees. For classification, the ensemble's output is the class label with the highest number of votes from all decision trees. The ensemble's output for regression is the average of the predicted values from all decision trees. The model's generalization and precision are enhanced by this ensemble method, which also helps prevent overfitting.

By combining the predictions from individual trees through voting or averaging, the Random Forest leverages the diversity and collective intelligence of the ensemble to provide robust and accurate predictions for both classification and regression tasks.

3.2. Elite Artificial Bee Colony Optimization

In 2005, Karaboga introduced the Artificial Bee Colony (ABC) optimization [25] method for continuous optimization, drawing inspiration from the foraging behaviour of honeybees in their colonies. The following sub-sections discuss the steps involved in Elite Artificial Bee Colony Optimization (EABC).

3.2.1. Random Population Initialization

The initial step involves generating a random population of T_m individuals within the interval $[-100,100]^c$, where c represents the total number of potential locations. This also serves as a gauge of the time required to code something from scratch. Eq.(10) mathematically describes the random population initialization process.

$$i_w^z = i^{min} + (i^{max} - i^{min}).rand() \quad (10)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where z belongs to the set $\{1,2,\dots,T_m\}$ and it represents the maximum value, w belongs to the set $\{1,2,\dots,c\}$ and it represents the minimum value, the $rand()$ function generates a pseudorandom scalar within the open interval $[0,1]$. Additionally, i^{max} and i^{min} represents the maximum and minimum values of the overall component.

3.2.2. Solution Update

The random population initialization step involves discretizing the continuous space that are available. Eq.(11) expresses how the solution is updated.

$$q_w^z = \lfloor mod(|i_w^z|,2) \rfloor \quad (11)$$

where, q_w^z is a binary variable that indicates the availability of building w for use. The modulus function $mod(i_1,i_2)$ returns the remainder obtained by dividing i_1 by i_2, while $\lfloor i_3 \rfloor$ represents the integer rounded down to the nearest whole number less than or equal to i_3 with all other arguments remaining unchanged.

All modifications are identified using the q_w^z values, from which the decision variables p_{sw} can be derived. In this scenario, s would be directed to the modification w that is globally closest to them among the available modifications. Therefore, if p_{sw} equals 1, it signifies that among all available facilities, facility w is the one closest to customer s. Conversely, if p_{sw} equals 0, it means that s is not directed to facility w.

3.2.3. Fitness Function

The fitness function for each z can be determined using the objective function, after the solution transformation step using Eq.(12).

$$fit_z = \begin{cases} 1/((1+g_z) \ if \ g_z \geq 0,) \\ 1+|g_z| \ otherwise. \end{cases} \quad (12)$$

where fit_z represents the fitness of individual z and g_z denotes the cost to the objective function for an individual i^z.

Algorithm 2: Fitness Function

Input:

Population: List of individuals, where each z is represented as a list of components.

$Objective_{Function}$: A function that takes an individual and returns the cost g_z to the objective function for that individual.

Output:

- A list of $fitness_{scores}$ containing the fitness values for each population member.

Procedure:

Initialize an empty list called fitness_scores to store the fitness values for each individual.

For each z in the population, do the following:

Calculate the cost g_z to the objective function for individual z using the provided $Objective_{Function}$.

Calculate the fitness value fit_z for individual z using Eq.(12):

if $g_z \geq 0$

$fit_z = 1 / (1 + g_z)$

else

$fit_z = 1 + |g_z|$

Append the calculated fit_z to the fitness_scores list.

End of the loop for individual z.

The $fitness_{scores}$ list contains the fitness values for all individuals.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Return the $fitness_{scores}$ list as the output.

3.2.4. Employed Bees

Employed bees will execute one of the fundamental phases of ABC to generate an entirely new individual r using Eq.(13).

$$r_w = i_w^z + (2 \times rand(\,) - 1).(i_w^z - \quad (13)$$

where w is a randomly selected integer from the set $\{1,2, \ldots, c\}$, distinct from the current person's z value, while r and a are random individuals chosen from the set $\{1,2, \ldots, T_m\}$. Only the most advantageous individual between r and i^z will persist, i.e., the superior individual at that moment will replace i^z.

3.2.5. Probabilistic Analysis

The probability values of an individual after the employed bees phase are calculated using Eq.(14). This probability value determines which individual will advance to the next level of exploitation.

$$m_z = \frac{fit_z}{\sum_{z_1=1}^{T_m} fit_{z_1}} \quad (14)$$

where m_z represents the probability of selecting individual z during the onlooker bees stage. z is an index from the set $\{1,2, \ldots, T_m\}$ representing individuals.

3.2.6. Observer Bees

In the second stage, a probability value is applied to the food sources discovered by the recruited bees, as shown in Eq.(10). Following Eq.(15), an onlooker bee will then exploit the chosen individual. A new potential member of the population, denoted as r, is generated. Similar to the earlier illustration with the worker bees, a competitive process occurs between r and i^{z3}.

$$r_w = i_w^{z2} + (2 * rand(\,) - 1).(i_w^{z2} - \quad (15)$$

Here, z_2 is a probability-weighted index representing a particular individual or food source selected from the set $\{1,2, \ldots, T_m\}$ based on probability weights. A random number greater than or equal to z_2 is denoted as a and is chosen from the set $\{1,2, \ldots, T_m\}$. All other values correspond to those defined in Eq. (9).

This procedure will continue until each observer bee discovers and utilizes a food source.

3.2.7. Bee Scouts phase

The current phase is deployed with the assistance of a threshold parameter, 'limit.' For simplicity, this research denotes the total number of times better solutions couldn't be found as fb^z for z. When a bee has been employed bee for a continuous cum consecutive period of fb^z greater than or equal to the 'limit,' it transitions to scout duty. Following Eq.(16), the current food source is immediately replaced with a newly generated random one. At this stage, only a single scout bee is randomly dispatched to explore a potential new food source. In other words, only the least effective worker bee is promoted to scout duty.

$$r_w = i^{min} + (i^{max} - i^{min}).rand(\,) \quad (16)$$

where w belongs to the set $\{1,2, \ldots, c\}$.

3.2.8. Restrictions on Handling Methods

In the context of the dynamic processes of EABC, it is essential to manage and adhere to certain constraints or restrictions on the components i^z, representing the potential solutions or individuals within the optimization problem. These constraints ensure that the search for the optimal solution remains within the specified bounds. The handling method for each component i^z, denoted as i_w^z, is designed to enforce these constraints effectively. It is a crucial part of the optimization process.

If the current value of i_w^z falls below the minimum permissible value i^{min}, it is adjusted using Eq.(17). A random value within the range [i^{min}, i^{max}] is generated

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and assigned to i_w^z. This ensures that the component remains within the lower bound.

$$i_w^z = i^{min} + (i^{max} - i^{min}).rand() \quad (17)$$

If the current value of i_w^z exceeds the maximum permissible value i^{max}, it is adjusted using Eq.(18). A random value within the same range [i^{min}, i^{max}] is generated and subtracted from i^{max}, ensuring the component stays within the upper bound.

$$i_w^z = i^{max} - (i^{max} - i^{min}).rand() \quad (18)$$

3.2.9 Crossovers Operator

The crossover operator is employed during the employed-bees phase of EABC to facilitate information sharing. Initially, individual i^z is replaced with i^z, and then a new perspective, r, is considered. Unlike the search technique of EABC, a randomly selected component of r can be updated using Eq.(19). It's evident that part w from a has replaced the corresponding part of z. Therefore, the new individual r obtains its w from the current individual i^z, while the remaining components originate from the previous individual a. This method effectively avoids altering the genetic makeup of the selected parent and can be considered a pure crossover operation.

$$r_w = i_w^a \quad (19)$$

where w is a randomly selected component, with w belonging to the set $\{1,2, \dots, c\}$. a is also a random number chosen from the set $\{1,2, \dots, T_m\}$, with the condition that a is not equal to z, where z is chosen from the set $\{1,2, \dots, T_m\}$. All other parameters remain the same as in Eq.(13).

3.2.10. Distribution Frequency

EABC presents a novel perturbation frequency strategy. One approach to modifying individuals is by adjusting the number of components the bee possesses. This number, denoted as Y, is a random integer within the range $[minY, maxY]$. Eq.(20) and Eq.(21) can be used to calculate both $minY$ and $maxY$.

$$minY = \left\lfloor \frac{c}{20} \right\rfloor + 1 \quad (20)$$

$$maxY = \sum \left\lfloor \frac{c}{20} \right\rfloor + 1 \quad (21)$$

where c represents the total number of possible infrastructures.

During the employed-bees phase of EABC, the Y dimensions of each individual, denoted as i^z, will be modified according to Eq.(18).

3.2.11. Frequency Pattern Perturbation

In the context of EABC, the algorithm implements a strategy to prevent the rapid aggregation of inferior individuals around superior ones during food source discovery. This is achieved by eliminating the initial probability of choosing a procedure.

(a). Initial Probability of Choosing a Procedure:

In typical optimization algorithms, an initial probability may be assigned to each bee (representing a potential solution) to choose a specific procedure or path to explore. However, in EABC, this initial probability is deliberately eliminated. This means that at the beginning of the optimization process, each bee does not have a predefined likelihood of choosing a particular procedure for exploring a food source.

(b). Preventing Rapid Aggregation of Inferior Individuals:

By removing the initial probability of procedure selection, EABC aims to mitigate the risk of rapid convergence of bees (individuals) towards a few promising solutions, which could be suboptimal or inferior. Without an initial bias towards specific

Journal of Theoretical and Applied Information Technology

15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

procedures, each bee has an equal chance to explore various food sources based on their characteristics and requirements.

(c). Ensuring Diverse Food Source Discovery:

This absence of an initial probability ensures diversity in exploring food sources. Each observing bee, regardless of its initial state, has an opportunity to discover a food source that best suits its requirements and characteristics. It encourages a more comprehensive search across the solution space, promoting the discovery of various food sources that may have different fitness values.

(d). Enhancing Solution Quality and Diversity:

By preventing the rapid aggregation of bees around inferior solutions, EABC enhances the overall quality and diversity of the solutions explored. It allows for a more thorough exploration of the solution space, increasing the likelihood of finding superior and diverse solutions well-suited to each observing bee's requirements.

3.2.12. Opposition-based Learning Method

Opposition-based Learning (OBL) is applied to mitigate the common issue of getting trapped in local optima in ABC. It's worth emphasizing that in the proposed EABC, every scout bee will discover a new food source, helping them avoid local traps more effectively. By combining the OBL approach with the distribution frequency strategy, a subset of components in i^{z3} for scout bee i^{z3} will be updated. According to Eq. (17), new components w can be generated randomly to replace the selected components.

$$r_w = i^{min} + i^{max} - i_w^{z3} \qquad (17)$$

If $w\,\omega\,\{1,2,\ldots,c\}$ then all other settings remain consistent with those described, representing an arbitrary index. EABC implements the OBL technique during scout bees with minimal adjustments to i^{z3}.

3.3. Fusion EABC and SRF

The fusion of Elite Artificial Bee Colony Optimization (EABC) and Synergy Random Forest (SRF) into "Elite Artificial Bee Colony Optimization-Based Synergy Random Forest (EABC-SRF)" represents a synergistic approach to machine learning and optimization. EABC-SRF harnesses the power of EABC's optimization capabilities to fine-tune the parameters of the SRF algorithm, enhancing its predictive performance. EABC optimizes SRF by iteratively adjusting its hyperparameters, allowing it to adapt to the intricacies of specific datasets. The fitness function of EABC evaluates SRF's performance, ensuring that the model evolves towards optimal configurations. EABC's bees, including employed, observer, and scout bees, explore the parameter space of SRF, fostering a diverse set of model configurations. SRF, on the other hand, brings its ensemble learning strength to the fusion. It leverages the synergy of multiple decision trees to make robust predictions. By integrating EABC's optimization process, SRF becomes more adaptable, accurate, and capable of handling complex datasets. The fusion's iterative nature ensures that the optimization continues until convergence, leading to highly tuned SRF models. EABC-SRF's automatic parameter tuning reduces the burden of manual hyperparameter selection and enhances the model's generalization ability.

EABC-SRF combines the optimization prowess of EABC with the ensemble learning capabilities of SRF, resulting in a powerful, adaptable, and accurate machine-learning framework. This fusion has the potential to significantly improve predictive modelling across various domains, making it a promising approach in machine learning and data science.

3.3.1. Advantages

- **Optimized Parameters:** EABC-SRF combines the power of optimization (EABC) with ensemble learning (SRF) to find the best parameter configurations for the random forest, leading to improved model performance.
- **Automatic Tuning:** The fusion allows for the automatic tuning of hyperparameters, reducing the need for manual parameter tuning.
- **Enhanced Generalization:** By optimizing SRF parameters, EABC-SRF aims to enhance the model's generalization ability and robustness, making it suitable for a wide range of datasets.
- **Diverse Ensemble**: SRF's ensemble approach ensures that the model leverages

the diversity of individual decision trees, leading to more accurate predictions.

Algorithm 3: EABC-SRF Algorithm

Step 1: Initialize parameters for EABC and SRF:
- **a.** Define EABC parameters (population size, maximum generations, bounds for parameters).
- **b.** Set up SRF parameters (number of decision trees, depth of trees).

Step 2: Initialize the population of parameter configurations for SRF using EABC:
- **a.** Generate a random population of parameter configurations.

Step 3: Evaluate the fitness of each parameter configuration:
- **a.** Train SRF with the given parameter configuration on a training dataset.
- **b.** Evaluate SRF's performance using a validation dataset.
- **c.** Assign fitness scores based on SRF's performance.

Step 4: Repeat the Steps 5 to 7 until convergence is attained:

Step 5: Employed Bees Phase (Parameter Modification):
- **a.** Select a subset of parameter configurations from the population.
- **b.** Modify the selected configurations to explore nearby regions in the parameter space.
- **c.** Evaluate the fitness of modified configurations.

Step 6: Onlooker Bees Phase (Probability-Based Selection):
- **a.** Select parameter configurations based on their fitness scores and probabilities.
- **b.** Modify the selected configurations.
- **c.** Evaluate the fitness of modified configurations.

Step 7: c. Scout Bees Phase (Exploration):
- **a.** Identify parameter configurations with low fitness scores (stagnant solutions).
- **b.** Replace these configurations with randomly generated ones.
- **c.** Evaluate the fitness of new configurations.

Step 8: Select the best parameter configuration based on fitness scores.

Step 9: Train the Synergy Random Forest (SRF) model using the selected parameter configuration on the training dataset.

Step 10: Make predictions on a test dataset using the trained SRF model.

Step 11: Evaluate the performance of the EABC-SRF model on the test.

Step 12: Output the final trained EABC-SRF model and its performance metrics.

4. ABOUT DATASET

The "Stock Tweets for Sentiment Analysis and Prediction" dataset is a comprehensive and invaluable resource for researchers and analysts in financial markets. Comprising an extensive collection of over 80,000 tweets, this dataset focuses on the top 25 most closely monitored stock tickers listed on Yahoo Finance. It spans a substantial time frame, covering the period from September 30, 2021, to September 30, 2022. This dataset's unique incorporation of real-time stock market price and volume data corresponding to each tweet and its associated stock sets this dataset apart. Each entry in this dataset offers essential information, including the precise date and time when the tweet was posted, the complete text of the tweet, the specific stock ticker name, and the corresponding company name. This dataset serves as a versatile tool for researchers and analysts, allowing them to harness its potential in various ways:

- **Sentiment Analysis:** By scrutinizing the sentiments expressed within these tweets, researchers can gain profound insights into public and investor sentiment regarding individual stocks. This analysis facilitates the detection of shifts and trends in market sentiment over the designated time frame.
- **Stock Price Prediction:** By integrating sentiment data with historical stock market data, analysts can develop predictive models capable of forecasting stock price movements. Such predictive capabilities offer a substantial advantage to investors and traders.
- **Exploration of Sentiment-Price Dynamics:** Researchers can explore the intricate connection between social media sentiment and subsequent stock price variations with this dataset. Uncovering

these patterns and relationships provides invaluable insights into the complex dynamics of financial markets.

This dataset, inspired by established sentiment analysis lexicons and existing stock market sentiment datasets, ensures its relevance and reliability. Whether you are a data scientist, financial analyst, or investor, this dataset opens doors to more profound insights into the intricate interplay between social media sentiment and the behaviour of financial markets, serving as a critical resource for advancing our understanding of stock market dynamics.

5. RESULTS AND DISCUSSION

5.1. Precision Analysis

Precision, in the context of this figure, quantifies the model's ability to make accurate positive predictions while minimizing false positives. It is calculated as the ratio of true positive predictions to the total positive predictions made by the model. Figure 1 provides the analysis of Precision results.

LSTM, a recurrent neural network (RNN) architecture, achieves a precision score of 47.77% in Figure 1. This value reflects LSTM's effectiveness in making positive predictions while maintaining reasonable precision. The architecture's focus on capturing sequential dependencies within data contributes to this result. SBRNM demonstrates a higher precision score of 60.13%. This suggests that SBRNM excels in distinguishing relevant positive instances from irrelevant ones. The bidirectional processing within the model allows it to consider both past and future contexts, enhancing its precision in positive predictions. EABC-SRF notably achieves the highest precision score of 88.72% among the models in Figure 1. This indicates EABC-SRF's exceptional ability to minimize false positive errors while providing accurate positive predictions. Utilizing elite artificial bee colony optimization and synergy within the random forest architecture likely significantly optimizes the model's parameters and enhances precision.

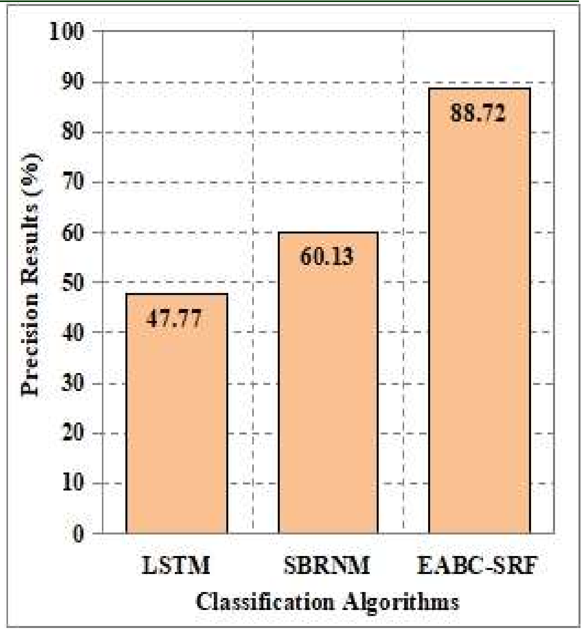

Figure 1. Precision

Figure 1 showcases the precision performance of different models, highlighting their varying abilities to make accurate positive predictions while avoiding false positives. LSTM, SBRNM, and EABC-SRF each contribute to precision differently based on their unique working mechanisms. Precision is a critical metric when selecting a model for applications where the cost of false positive errors is high, ensuring that positive predictions are highly reliable and accurate.

5.2. Recall Analysis

Figure 2 presents the recall performance metric vital for evaluating the effectiveness of three distinct models: LSTM, SBRNM, and EABC-SRF. Recall assesses the models' ability to correctly identify and capture a high percentage of relevant positive instances within the dataset. Recall is crucial to minimize false negatives, ensuring that as few positive instances as possible are missed.

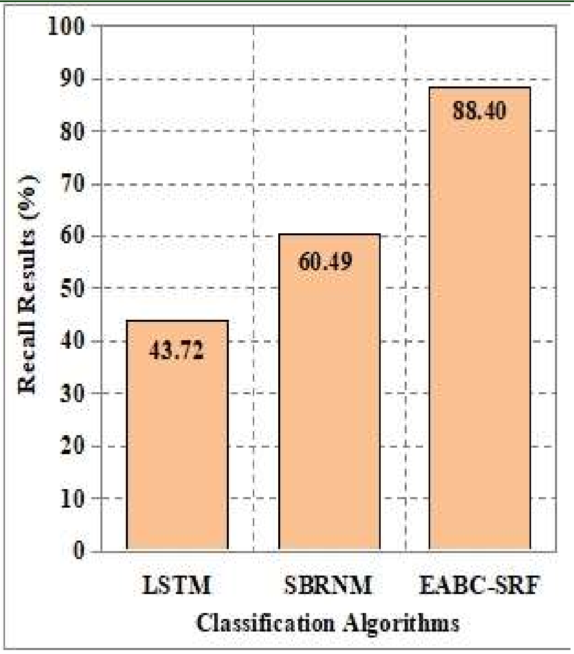

Figure 2. Recall

LSTM, a recurrent neural network (RNN) architecture, achieves a recall score of 43.72% in Figure 2. This score reflects LSTM's ability to capture a substantial proportion of positive instances within the data. LSTM's strength lies in its capacity to recognize patterns and dependencies within sequences, contributing to its recall performance. SBRNM demonstrates a higher recall score of 60.49%. This indicates that SBRNM excels in correctly identifying and capturing a significant portion of the relevant positive instances. The bidirectional processing employed by SBRNM allows it to consider both past and future context, making it adept at recognizing nuanced patterns in the data. EABC-SRF attains a recall score of 88.40%, the highest among the models in Figure 2. This underscores EABC-SRF's exceptional ability to identify and capture the vast majority of actual positive instances. Integrating elite artificial bee colony optimization and synergy within the random forest architecture is pivotal in optimizing the model's parameters for recall-oriented tasks.

Figure 2 highlights the recall performance of different models, emphasizing their varying abilities to effectively identify and capture actual positive instances. While LSTM, SBRNM, and EABC-SRF contribute to recall differently, their distinct working mechanisms result in different performance levels in this crucial metric. Recall is particularly valuable in applications where missing positive cases, such as medical diagnosis or fraud detection, can have significant consequences.

5.3. Classification Accuracy Analysis

Figure 3 is a fundamental evaluation measure for three distinct models: LSTM, SBRNM, and EABC-SRF. Classification accuracy assesses the overall correctness of a model's predictions across all classes within a dataset. This metric is paramount in various machine learning applications, reflecting the model's ability to provide accurate predictions for both positive and negative instances.

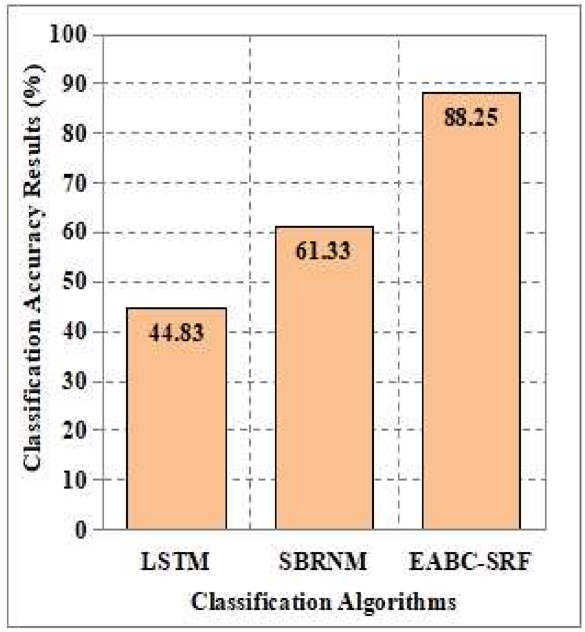

Figure 3. Classification Accuracy

LSTM, a recurrent neural network (RNN) architecture, attains an accuracy score of 44.83% in Figure 3. This score indicates LSTM's competence in making accurate predictions across all classes, although it may face challenges when dealing with complex or multi-modal data. LSTM's strength lies in its capacity to capture sequential dependencies within data, contributing to its accuracy. SBRNM achieves a higher accuracy score of 61.33%. This highlights SBRNM's effectiveness in accurately classifying instances across various classes within the dataset. The bidirectional processing within SBRNM allows it to consider past and future contexts, enhancing its ability to recognize diverse patterns in the data. EABC-SRF stands out with an accuracy score of 88.25%, the highest among the models in Figure 3. This underscores EABC-SRF's exceptional capability to provide accurate predictions across multiple classes. Integrating elite artificial bee colony

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

optimization and synergy within the random forest architecture significantly optimizes the model's parameters for high overall accuracy.

Figure 3 highlights the classification accuracy performance of different models, emphasizing their varying abilities to provide accurate predictions across all classes within the dataset. While LSTM, SBRNM, and EABC-SRF contribute to accuracy differently, their distinct working mechanisms result in different performance levels in this crucial metric. Classification accuracy is critical when selecting a model for tasks where overall correctness in classifying instances is essential, such as image recognition or text categorization.

5.4. F-Measure Analysis

Figure 4 is a comprehensive measure of model performance for three distinctive models: LSTM, SBRNM, and EABC-SRF. The F-Measure evaluates a model's precision and recall, striking a harmonious trade-off between these critical aspects. This metric is particularly valuable when maintaining a balance between minimizing false positives (precision) and false negatives (recall) is essential, ensuring that optimistic predictions are accurate and comprehensive.

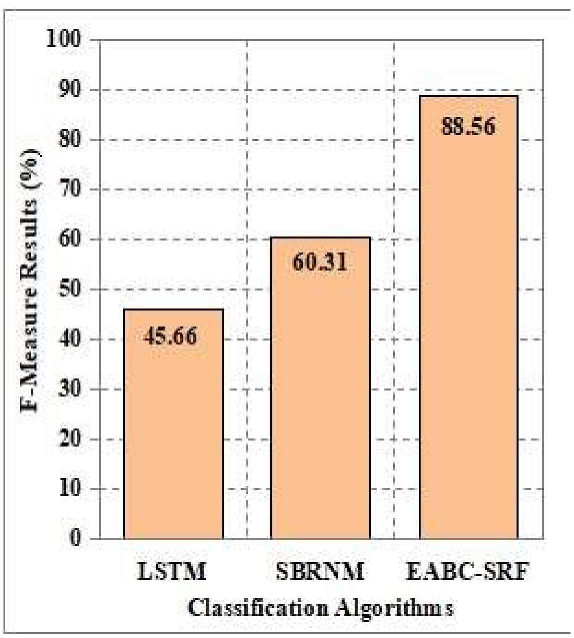

Figure 4. F-Measure

LSTM yields an F-Measure of 45.66% in Figure 4. This indicates LSTM's ability to balance precision and recall, capturing sequential dependencies in data and achieving a reasonable balance between false positives and false negatives. SBRNM achieves a slightly higher F-measure of 60.31%. This suggests that SBRNM excels in finding a balance between precision and recall, capturing relevant patterns in the data while minimizing false positives and negatives. The bidirectional processing within SBRNM enhances its performance in achieving this harmony. EABC-SRF stands out with an F-Measure of 88.56%, the highest among the models in Figure 4. This highlights EABC-SRF's exceptional capability to optimize parameters effectively, striking an excellent balance between precision and recall. Integrating elite artificial bee colony optimization and synergy within the random forest architecture plays a pivotal role in achieving this balance.

Figure 4 visualizes the F-Measure performance of different models, emphasizing their varying abilities to strike a harmonious balance between precision and recall. While LSTM, SBRNM, and EABC-SRF each contribute to the F-Measure differently, their distinct working mechanisms result in different performance levels in this crucial metric. The F-Measure is particularly valuable in scenarios where it is crucial to balance minimizing false positives and negatives, such as medical diagnostics or anomaly detection, ensuring that positive predictions are accurate and comprehensive.

6. CONCLUSION

This research underscores the significance of Twitter as a real-time information hub within the intricate realm of the stock market. In a landscape influenced by a multitude of factors, ranging from economic indicators to geopolitical events, the role of social sentiment, as manifested in stock tweets, emerges as a dynamic and influential force with the capacity to profoundly impact investor decisions and market trends. As demonstrated in this research, the categorization of stock tweets based on sentiment proves invaluable for comprehending investor sentiment and prognosticating market dynamics. The proposed optimization-based classification algorithm, namely Elite Artificial Bee Colony Optimization-Based Synergy Random Forest (EABC-SRF), represents a notable leap forward in sentiment analysis. By synergizing the capabilities of Artificial Bee Colony Optimization (ABC) and Synergy Random Forest (SRF), EABC-SRF adeptly addresses the inherent complexity and noise characteristic of stock tweets, thereby elevating the precision and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

transparency of sentiment classification. EABC-SRF's feature selection mechanism, facilitated by elite artificial bees, and its seamless integration into the SRF framework underscore the potential of this novel approach in deciphering intricate social sentiment data. The reliance on the "Stock Tweets for Sentiment Analysis and Prediction" dataset as the cornerstone for developing and assessing EABC-SRF underscores the fundamental role of robust data sources in advancing such models. The outstanding performance of EABC-SRF, as substantiated by experimental outcomes, reaffirms its efficacy in aiding investors and traders by furnishing perceptive sentiment analysis to facilitate more discerning decision-making within the ever-evolving stock market milieu. This research underscores the pivotal role of cutting-edge machine learning techniques in distilling invaluable insights from the ever-expansive domain of social media data, thereby contributing to more astute market analysis and prediction. Performance of the current research can be enhanced with optimization strategies in future.

REFERENCES

[1] H. Ni, S. Wang, and P. Cheng, "A hybrid approach for stock trend prediction based on tweets embedding and historical prices," *World Wide Web*, vol. 24, no. 3, pp. 849–868, 2021, doi: 10.1007/s11280-021-00880-9.

[2] A. Fekrazad, S. M. Harun, and N. Sardar, "Social media sentiment and the stock market," *J. Econ. Financ.*, vol. 46, no. 2, pp. 397–419, 2022, doi: 10.1007/s12197-022-09575-x.

[3] S. García-Méndez, F. de Arriba-Pérez, A. Barros-Vila, and F. J. González-Castaño, "Targeted aspect-based emotion analysis to detect opportunities and precaution in financial Twitter messages," *Expert Syst. Appl.*, vol. 218, p. 119611, 2023, doi: 10.1016/j.eswa.2023.119611.

[4] A. Abdurrahim, L. Syafa'ah, and M. Lestandy, "Sentiment analysis of Covid-19 vaccine tweets utilizing Naïve Bayes," in *AIP Conference Proceedings*, 2022. doi: 10.1063/5.0094607.

[5] Y. Nishimura and B. Sun, "President's Tweets, US-China economic conflict and stock market Volatility: Evidence from China and G5 countries," *North Am. J. Econ. Financ.*, vol. 58, p. 101506, 2021, doi: 10.1016/j.najef.2021.101506.

[6] S. Bahri, P. Bahri, and S. Lal, "A Novel approach of Sentiment Classification using Emoticons," in *Procedia Computer Science*, Elsevier B.V., Jan. 2018, pp. 669–678. doi: 10.1016/j.procs.2018.05.067.

[7] P. K. Roy, A. Singh, A. K. Tripathy, and T. K. Das, "Cyberbullying detection: an ensemble learning approach," *Int. J. Comput. Sci. Eng.*, vol. 25, no. 3, pp. 315–324, 2022, doi: 10.1504/IJCSE.2022.123121.

[8] S. Fong, R. Wong, and A. V. Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data," *IEEE Trans. Serv. Comput.*, vol. 9, no. 1, pp. 33–45, Jan. 2016, doi: 10.1109/TSC.2015.2439695.

[9] M. I. Alghamdi, "Single Valued Neutrosophic Sets with Optimal Support Vector Machine for Sentiment Analysis," *Int. J. Neutrosophic Sci.*, vol. 18, no. 3, pp. 227–237, 2022, doi: 10.54216/IJNS.1803020.

[10] G. Yildirim, "A novel grid-based many-objective swarm intelligence approach for sentiment analysis in social media," *Neurocomputing*, vol. 503, pp. 173–188, 2022, doi: 10.1016/j.neucom.2022.06.092.

[11] K. A. Hadi, R. Lasri, and A. El Abderrahmani, "An efficient approach for sentiment analysis in a big data environment," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4, pp. 263–266, 2019, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067872768&partnerID=40&md5=da8474ea6d562432f7d00af5293b9485

[12] A. Senthilkumar, J. Ramkumar, M. Lingaraj, D. Jayaraj, and B. Sureshkumar, "Minimizing Energy Consumption in Vehicular Sensor Networks Using Relentless Particle Swarm Optimization Routing," *Int. J. Comput. Networks Appl.*, vol. 10, no. 2, pp. 217–230, 2023, doi: 10.22247/ijcna/2023/220737.

[13] J. Ramkumar, S. S. Dinakaran, M. Lingaraj, S. Boopalan, and B. Narasimhan, "IoT-Based Kalman Filtering and Particle Swarm Optimization for Detecting Skin Lesion," in *Lecture Notes in Electrical Engineering*, K. Murari, N. Prasad Padhy, and S. Kamalasadan, Eds., Singapore: Springer Nature Singapore, 2023, pp. 17–27. doi: 10.1007/978-981-19-8353-5_2.

[14] D. Jayaraj, J. Ramkumar, M. Lingaraj, and B. Sureshkumar, "AFSORP: Adaptive Fish Swarm Optimization-Based Routing Protocol for Mobility Enabled Wireless Sensor Network," *Int. J. Comput. Networks Appl.*,

vol. 10, no. 1, pp. 119–129, Jan. 2023, doi: 10.22247/ijcna/2023/218516.

[15] J. Ramkumar, C. Kumuthini, B. Narasimhan, and S. Boopalan, "Energy Consumption Minimization in Cognitive Radio Mobile Ad-Hoc Networks using Enriched Ad-hoc On-demand Distance Vector Protocol," *2022 Int. Conf. Adv. Comput. Technol. Appl. ICACTA 2022*, pp. 1–6, Mar. 2022, doi: 10.1109/ICACTA54488.2022.9752899.

[16] P. Menakadevi and J. Ramkumar, "Robust Optimization Based Extreme Learning Machine for Sentiment Analysis in Big Data," *2022 Int. Conf. Adv. Comput. Technol. Appl. ICACTA 2022*, pp. 1–5, Mar. 2022, doi: 10.1109/ICACTA54488.2022.9753203.

[17] J. Ramkumar, R. Vadivel, and B. Narasimhan, "Constrained Cuckoo Search Optimization Based Protocol for Routing in Cloud Network," *Int. J. Comput. Networks Appl.*, vol. 8, no. 6, pp. 795–803, 2021, doi: 10.22247/ijcna/2021/210727.

[18] J. Ramkumar and R. Vadivel, "Multi-Adaptive Routing Protocol for Internet of Things based Ad-hoc Networks," *Wirel. Pers. Commun.*, vol. 120, no. 2, pp. 887–909, Apr. 2021, doi: 10.1007/s11277-021-08495-z.

[19] J. Ramkumar and R. Vadivel, "Whale optimization routing protocol for minimizing energy consumption in cognitive radio wireless sensor network," *Int. J. Comput. Networks Appl.*, vol. 8, no. 4, pp. 455–464, 2021, doi: 10.22247/ijcna/2021/209711.

[20] R. Jaganathan and R. Vadivel, "Intelligent Fish Swarm Inspired Protocol (IFSIP) for Dynamic Ideal Routing in Cognitive Radio Ad-Hoc Networks," *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 1063–1074, 2021, doi: 10.12785/ijcds/100196.

[21] J. Ramkumar and R. Vadivel, "Improved Wolf prey inspired protocol for routing in cognitive radio Ad Hoc networks," *Int. J. Comput. Networks Appl.*, vol. 7, no. 5, pp. 126–136, 2020, doi: 10.22247/ijcna/2020/202977.

[22] R. Vadivel and J. Ramkumar, "QoS-enabled improved cuckoo search-inspired protocol (ICSIP) for IoT-based healthcare applications," *Inc. Internet Things Healthc. Appl. Wearable Devices*, pp. 109–121, 2019, doi: 10.4018/978-1-7998-1090-2.ch006.

[23] R. Jaganathan and V. Ramasamy, "Performance modeling of bio-inspired routing protocols in Cognitive Radio Ad Hoc Network to reduce end-to-end delay," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 221–231, 2019, doi: 10.22266/IJIES2019.0228.22.

[24] J. Ramkumar and R. Vadivel, "Improved frog leap inspired protocol (IFLIP) – for routing in cognitive radio ad hoc networks (CRAHN)," *World J. Eng.*, vol. 15, no. 2, pp. 306–311, 2018, doi: 10.1108/WJE-08-2017-0260.

[25] J. Ramkumar and R. Vadivel, "CSIP—cuckoo search inspired protocol for routing in cognitive radio ad hoc networks," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2017, pp. 145–153. doi: 10.1007/978-981-10-3874-7_14.

[26] L. Mani, S. Arumugam, and R. Jaganathan, "Performance Enhancement of Wireless Sensor Network Using Feisty Particle Swarm Optimization Protocol," *ACM Int. Conf. Proceeding Ser.*, pp. 1–5, Dec. 2022, doi: 10.1145/3590837.3590907.

[27] C. Guan, W. Liu, and J. Y. C. Cheng, "Using Social Media to Predict the Stock Market Crash and Rebound amid the Pandemic: The Digital 'Haves' and 'Have-mores,'" *Ann. Data Sci.*, vol. 9, no. 1, pp. 5–31, 2022, doi: 10.1007/s40745-021-00353-w.

[28] Y. Zhang, M. Shirakawa, Y. Wang, Z. Li, and T. Hara, "Twitter-aided decision making: a review of recent developments," *Appl. Intell.*, vol. 52, no. 12, pp. 13839–13854, 2022, doi: 10.1007/s10489-022-03241-9.

[29] S. Smith and A. O'Hare, "Comparing traditional news and social media with stock price movements; which comes first, the news or the price change?," *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00591-6.

[30] K. Kim, S. Y. T. Lee, and R. J. Kauffman, "Social informedness and investor sentiment in the GameStop short squeeze," *Electron. Mark.*, vol. 33, no. 1, p. 23, 2023, doi: 10.1007/s12525-023-00632-9.

[31] R. Fan, O. Talavera, and V. Tran, "Social media, political uncertainty, and stock markets," *Rev. Quant. Financ. Account.*, vol. 55, no. 3, pp. 1137–1153, 2020, doi: 10.1007/s11156-020-00870-4.

[32] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 3007–3057, 2020, doi: 10.1007/s10462-019-09754-z.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[33] S. M. Sabharwal and N. Aggrawal, "A Survey on Information Diffusion over Social Network with the Application on Stock Market and its Future Prospects," *Wirel. Pers. Commun.*, vol. 130, no. 4, pp. 2981–3007, 2023, doi: 10.1007/s11277-023-10412-5.

[34] A. Nekrasov, S. H. Teoh, and S. Wu, "Visuals and attention to earnings news on twitter," *Rev. Account. Stud.*, vol. 27, no. 4, pp. 1233–1275, 2022, doi: 10.1007/s11142-021-09630-8.

[35] J. U. N. Vogel and F. Xie, "Do(n't) believe everything you hear about disclosure: Twitter and the voluntary disclosure effect," *Financ. Mark. Portf. Manag.*, vol. 37, no. 2, pp. 161–189, 2023, doi: 10.1007/s11408-022-00420-z.

[36] Z. McGurk, A. Nowak, and J. C. Hall, "Stock returns and investor sentiment: textual analysis and social media," *J. Econ. Financ.*, vol. 44, no. 3, pp. 458–485, 2020, doi: 10.1007/s12197-019-09494-4.

[37] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," *PeerJ Comput. Sci.*, vol. 7, pp. 1–21, 2021, doi: 10.7717/peerj-cs.476.

[38] T. Pattewar and D. Jain, "Stock prediction analysis by customers opinion in Twitter data using an optimized intelligent model," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-022-00979-5.