SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

A Hybrid Framework for Stock-Market Prediction Using COVID-19 Tweets through Nature-Inspired Algorithms and Machine Learning

Muhammad Nauman Khan[1], Saman Iftikhar[2*], Daniah Al-Madani[2], Umer Zaheer[1], Danish khan[3], Kiran Fatima[4], Ammar Saeed[3]

1 Faculty of Business Studies, Arab Open University, Saudi Arabia. Emails: m.nauman@arabou.edu.sa, umerzaheerbabar@gmail.com
2 Faculty of Computer Studies, Arab Open University, Saudi Arabia. Emails: s.iftikhar@arabou.edu.sa, d.almadani@arabou.edu.sa.
3 COMSATS University Islamabad, Wah Campus, Pakistan. Emails: danish56566@gmail.com, ammarsaeed1997@gmail.com
4 Technical and Further Education, TAFE, New South Wales, Australia. Email: kiran.fatima4@tafensw.edu.au.

ARTICLE INFO

Article history:
Received 15 April 2025
Received in revised form 25 May 2025
Accepted 27 June 2025
Available online 01 December 2025

Keywords:
Evolutionary Computing, Stock Exchange Prediction, Covid-19, Evolutionary Algorithm, Particle Swarm Optimization, and Genetic Algorithm.

ABSTRACT

The COVID-19 outbreak generated severe disturbances across international financial systems, intensifying volatility within equity markets. Stock market behaviour during this period exhibited pronounced instability, shaped by broader economic conditions as well as collective investor sentiment. This study investigates the effect of COVID-19 related discourse on Twitter on market behaviour and proposes a hybrid modelling framework aimed at enhancing the accuracy of stock price movement forecasts. The proposed framework incorporates nature inspired optimisation techniques, including genetic algorithms, Harris hawk's optimisation, particle swarm optimisation, and differential evolution. The methodological focus lies in embedding social media sentiment into predictive models to uncover latent relationships that are typically overlooked by conventional statistical approaches. Empirical results demonstrate that all hybrid configurations consistently surpassed traditional forecasting methods. Among them, the support vector regression model optimised using Harris hawk's optimisation achieved superior performance, recording an MSE of 0.00014, an RMSE of 0.00214, and an MAE of 0.00170. These findings underscore the substantial influence of public sentiment on financial market dynamics during periods of global disruption and highlight the effectiveness of hybrid predictive approaches. Such models offer valuable support for stock market forecasting and provide actionable insights for decision makers operating under heightened economic uncertainty.

1. Introduction

The Saudi stock market represents a central pillar of the national economy and attracts sustained interest from both domestic and international stakeholders. Market participants, including investors, financial analysts, and policy makers, have strong incentives to understand its performance and the factors that shape its behaviour. Traditionally, empirical investigations have relied on structured financial indicators such as share prices, trading volumes, and corporate earnings. However, recent

* *Corresponding author.*
E-mail address: m.nauman@arabou.edu.sa

https://doi.org/10.31081/dmame8220251600

research highlights the value of complementing these conventional data sources with unstructured information, including news content and social media platforms such as Twitter, to achieve a more comprehensive understanding of market dynamics [3].

Unstructured data evolves rapidly and reflects the emotions, perceptions, and behavioural responses of market participants, particularly investors, which are not readily observable through standard financial metrics. For example, following the emergence of the Omicron variant, online discussions were dominated by negative sentiment, underscoring the importance of monitoring public emotions during the spread of new diseases and adopting effective communication strategies in response [21]. The outbreak of COVID-19 triggered profound transformations in business operations. Telework practices were redefined, and the organisational and employee level responses to this shift produced notable effects within the Saudi Arabian context. Analysing public sentiment towards teleworking during this period enabled organisations to better understand how this working arrangement could be implemented to maximise organisational benefits [5].

In the education sector, the pandemic led to widespread institutional closures, accelerating the adoption of e learning systems. Public perceptions of this transition were assessed through Twitter based analyses, revealing variations in attitudes towards online education [7]. Beyond these domains, social media platforms played a critical role in the dissemination of information throughout the pandemic. While these platforms facilitated open expression of opinions, they also contributed to the rapid spread of misinformation. Prior studies emphasise that social media data must be carefully processed to extract meaningful insights and to address challenges associated with misleading or false information [4]. Despite the relevance of these developments, a notable gap persists in incorporating such insights into predictive models of financial markets, particularly within the Saudi stock market context. Many existing forecasting approaches fail to adequately account for public sentiment derived from unstructured data, which may result in less reliable guidance for investors. In addition, conventional models frequently rely on linear assumptions, limiting their ability to capture the complex and potentially nonlinear relationships between collective sentiment and stock market movements.

To address these limitations, the proposed study introduces a novel hybrid framework that integrates evolutionary optimisation techniques with advanced sentiment analysis methods. The primary objective is to construct a comprehensive data set that combines historical numerical indicators of stock market performance with unstructured textual information obtained from financial news articles and microblogging platforms related to the Saudi stock market. By applying natural language processing techniques to determine sentiment polarity within the unstructured data, the framework aims to identify latent patterns and non-linear associations that are often overlooked by traditional econometric models that depend exclusively on historical price information. Furthermore, the proposed model is designed as a sophisticated system capable of enhancing the accuracy of stock market forecasting. In its initial stage, the framework addresses financial data through an automated feature selection process. Nature inspired optimisation algorithms, including genetic algorithms, Harris hawks optimisation, particle swarm optimisation, and differential evolution, are employed to efficiently identify the most informative features from high dimensional data sets. The careful selection of relevant inputs is critical for improving analytical robustness. The refined data are subsequently utilised within advanced machine learning and deep learning models to estimate stock market behaviour. Model performance is evaluated using

standard error metrics, namely mean squared error, root mean squared error, and mean absolute error. Through the execution of this study, several key objectives are intended to be achieved:

- To enhance the accuracy of stock market trend forecasting by incorporating sentimental information derived from news reports and social networking platforms.
- To identify and analyse the principal factors influencing the Saudi stock market through sentiment based analytical techniques.
- To demonstrate the effectiveness of evolutionary algorithms in strengthening predictive modelling performance.
- To highlight the superiority of evolutionary algorithms over traditional approaches in improving prediction accuracy through enhanced optimisation capabilities.

The outcomes of this research are expected to contribute meaningfully to the advancement of computational finance by enabling a deeper understanding of the Saudi stock market and supporting more informed investment decision making. The study aims to address the shortcomings of earlier models that did not integrate real time sentiment driven data, an omission that has become increasingly critical in the context of the COVID-19 pandemic. By doing so, the research seeks to overcome specific challenges associated with the analysis of unstructured data and to improve the reliability of stock market performance forecasting. The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of prior studies related to stock market prediction and sentiment analysis. Section 3 outlines the proposed methodology, including data acquisition procedures, feature selection strategies, and the development of the hybrid predictive framework. A detailed case study is presented in Section 4. Section 5 discusses empirical results and compares the performance of the proposed models with existing approaches. Finally, Section 6 summarises the principal findings and key contributions of the study.

2. Related Work

The literature demonstrates growing interest in stock market forecasting through the exploitation of Twitter data, emphasising its role in shaping and explaining market trends. Prior studies have extensively reviewed the use of Twitter based information in financial analysis, particularly focusing on the application of sentiment analysis and natural language processing techniques to examine how collective emotions expressed in tweets influence stock market behaviour. These approaches have consistently shown that market movements are not solely driven by numerical indicators but are also strongly affected by public sentiment extracted from social media streams.

Empirical evidence indicates that sentiment indicators derived from Twitter can anticipate stock market behaviour with notable accuracy. Sentiment analysis techniques extract subjective signals from Twitter data and categorise them into positive, negative, or neutral classes using a range of approaches, including lexicon-based methods and deep learning models. Such sentiment signals provide insights into investor psychology and market mood, offering advantages over traditional data sources due to their real time nature and representation of collective perceptions. Within the Saudi Arabian context, Arabic sentiment analysis has been applied to assess public reactions to COVID-19, revealing pandemic related emotional patterns and highlighting the relevance of social media data for understanding societal and market level responses [6].

Several studies have explicitly examined the relationship between Twitter activity and stock market movements during the COVID-19 period. One investigation developed the LOGCOVID19 framework using the autoregressive distributed lag approach to explore the association between

Twitter discussions and fluctuations in the Saudi stock market. Using data from the Tadawul exchange covering March to May 2020, the study analysed the relationship between stock index movements and COVID-19 related tweet volumes, demonstrating that public opinion significantly influenced investor activity during the pandemic [13]. In a related strand of research, machine learning and deep learning models were evaluated for predicting stock price trends in the Egyptian real estate sector over the period 2013 to 2022. The findings showed that ensemble learning techniques and recurrent neural networks achieved superior predictive accuracy, underscoring their ability to capture complex patterns in financial time series and support informed investment decisions [14].

Comparative modelling frameworks have also been proposed to assess market behaviour under extreme global conditions. One study benchmarked econometric and artificial intelligence models, including hybrid configurations that combined neural networks with optimisation techniques, to forecast Brent crude oil prices during COVID-19 and the Russia Ukraine conflict. The results indicated that optimisation enhanced neural models substantially reduced prediction errors, highlighting the benefits of integrating econometric reasoning with machine learning under volatile market conditions [22]. Similarly, explainable machine learning frameworks have been applied to address challenges associated with high dimensionality, multicollinearity, and interpretability in financial prediction tasks. A large-scale study employing optimisation based hyperparameter tuning demonstrated that gradient boosting models could explain a substantial proportion of variance in corporate financialisation, while interpretability tools revealed non-linear effects of multiple economic and governance related variables [27].

Advanced hybrid architectures that integrate evolutionary optimisation with deep sequential models have further strengthened predictive reliability. An evolutionary bi directional long short-term memory framework combined with sentiment classification techniques demonstrated high precision and robustness in volatile stock market environments, particularly during crisis periods such as COVID-19 [24]. Ensemble based approaches have also been shown to enhance forecasting accuracy in energy markets, where residual driven learning frameworks significantly improved performance during periods of heightened uncertainty and suppressed data availability [19]. In foreign exchange markets, ensemble deep learning models integrating bidirectional recurrent networks and regression techniques outperformed both traditional machine learning and standalone deep learning approaches, reflecting the increased volatility observed during the pandemic [2].

Explainable machine learning has been applied to evaluate the effects of government intervention measures on sectoral stock market performance during COVID-19. Using interpretable boosting models, studies have demonstrated that policy interventions exert heterogeneous and asymmetric impacts on stock returns and volatility across different sectors, reinforcing the importance of transparent predictive frameworks for policy relevant financial analysis [29]. Bio inspired supervised learning algorithms have also emerged as competitive alternatives to classical statistical and deep learning methods. Comparative evaluations using national stock indices have shown that such models achieve high predictive accuracy and adaptability when handling complex financial systems [16]. Sentiment driven forecasting has further been extended to sector specific applications. Hybrid frameworks combining composite sentiment indices with convolutional neural networks have delivered improved stock price predictions in the halal tourism sector, while also revealing aspects of irrational investor behaviour through explainable artificial intelligence

mechanisms [1]. Additional research has proposed optimisation driven sentiment analysis frameworks for decoding public mood from COVID-19 related tweets, demonstrating the effectiveness of combining natural language processing with nature inspired optimisation strategies [26]. Multi objective optimisation techniques integrated with Transformer based architectures have also been validated across benchmark problems and real-world stock datasets, consistently outperforming classical neural and optimisation baselines in terms of accuracy and stability [28].

Beyond finance specific applications, hybrid deep learning and evolutionary optimisation frameworks have been successfully applied in other complex domains, including medical image classification and segmentation, demonstrating the general effectiveness of optimisation enhanced learning systems [9]. In the financial domain, attention-based neuro fuzzy systems optimised using genetic algorithms have shown superior performance in stock market prediction tasks, confirming the interdependence of optimisation, attention mechanisms, and fuzzy logic in achieving robust forecasts [17]. Novel optimisation algorithms that improve exploration and exploitation balance have also been benchmarked extensively and applied to real world engineering and security problems, further validating their robustness and convergence efficiency [15].

Within the area of Arabic sentiment analysis, machine learning frameworks based on word embedding and ensemble learning have demonstrated substantial improvements in classification performance, particularly when combined with data balancing techniques [3]. Multi class sentiment classification models applied to COVID-19 related tweets have achieved high accuracy rates, providing insights into public emotional responses and confirming the effectiveness of advanced sentiment analysis methodologies [18]. Lexicon based analyses have additionally been used to assess shifts in public opinion before and after the pandemic, revealing significant sentiment changes and potential links to national stock index performance [30]. Several studies have extended optimisation enhanced deep learning frameworks to pandemic related diagnosis, forecasting, and misinformation detection. Hybrid systems combining convolutional neural networks with data augmentation techniques have achieved very high accuracy in COVID-19 diagnosis and prognosis tasks [10], while optimisation assisted deep learning models have further improved detection performance using medical imaging data [10]. Enhanced swarm intelligence algorithms have demonstrated superiority over existing methods on benchmark functions and real-world control applications [25]. Research on automated text similarity and assessment systems has also proposed novel frameworks that outperform existing symbolic computation tools [11].

In cryptocurrency markets, hybrid decomposition and recurrent neural network models optimised through metaheuristic algorithms have shown strong predictive capability, identifying key macroeconomic and digital asset features that influence price volatility [23]. Analytical frameworks integrating forecasting models with portfolio optimisation techniques have been proposed to help investors manage risk and exploit opportunities during periods of market disruption, such as the COVID-19 crisis [20]. Optimisation based feature selection has also been applied to fake news detection, improving classification accuracy and supporting efforts to mitigate misinformation during the pandemic [31]. Finally, hybrid extreme learning machine models optimised using multiple evolutionary algorithms have demonstrated high predictive accuracy across different industrial sectors, while neural optimisation-based models have achieved improved performance in long term stock index forecasting tasks [8; 12].

Collectively, these studies underline the growing importance of hybrid predictive frameworks that integrate sentiment analysis, evolutionary optimisation, and advanced machine learning

techniques. They also reveal a clear trajectory towards explainable, optimisation enhanced models capable of capturing complex, non-linear relationships in financial markets, particularly under conditions of uncertainty and systemic shock such as the COVID-19 pandemic.

3. Proposed Methodology

Conventional statistical approaches, including correlation analysis, regression techniques, and hypothesis testing, are effective in identifying linear relationships and stable trends under normal market conditions. However, during crisis periods following the COVID-19 pandemic, financial markets are driven by highly non-linear, volatile, and continuously evolving mechanisms. Stock market indices are shaped not only by historical price information but also by public perceptions and emotions, which are commonly expressed through unstructured data such as tweets. Traditional statistical techniques are inherently limited in capturing these complex interactions and therefore struggle to provide reliable insights under such circumstances. Machine learning and deep learning techniques are better suited to address these challenges, as they are designed to detect hidden structures and complex dependencies within large and heterogeneous datasets. Models such as support vector regression and long short-term memory networks can learn adaptively from both numerical and textual inputs, enabling them to respond to evolving patterns that emerge from shifts in collective sentiment and market behaviour.

Feature selection represents a critical stage when integrating social media data with financial variables. Meta heuristic optimisation techniques, including genetic algorithms, particle swarm optimisation, Harris hawk's optimisation, and differential evolution, provide effective mechanisms for noise reduction, identification of informative features, and enhancement of learning efficiency within ML and DL frameworks. Unlike deterministic feature selection methods, meta heuristic approaches emulate natural search processes, allowing them to escape local optima and better explore high dimensional and non-linear search spaces. As a result, hybrid models that combine ML or DL techniques with evolutionary optimisation exhibit greater robustness and improved predictive accuracy when dealing with complex financial data.

To examine the impact of the COVID-19 pandemic on stock market indices, this study proposes a hybrid framework based on evolutionary computation. Tweets related to pandemic developments and stock market activities are collected through the Twitter application programming interface using carefully selected keywords. Corresponding stock index data, including major indices such as the S&P 500 and the Dow Jones, are retrieved from Yahoo Finance. Both data sources are temporally aligned to ensure consistency and comparability. The pre-processing phase plays a crucial role in eliminating noise and ambiguity from the raw data. Non-informative elements such as hyperlinks, punctuation marks, and stop words are removed, followed by tokenisation to convert textual content into individual terms. Lemmatization and stemming are subsequently applied to normalise words to their root forms. For numerical variables, min max normalisation is employed to ensure uniform scaling across features. Subsequently, textual and emotional features are extracted using natural language processing techniques, while stock market features are derived from daily closing price data. Evolutionary algorithms, namely genetic algorithms, particle swarm optimisation, Harris hawks optimisation, and differential evolution, are applied sequentially to enhance feature selection and reduce computational complexity. This process filters out redundant or irrelevant variables and ensures that only the most informative features are retained, thereby improving model performance.

Following optimisation, the dataset is divided into training and testing subsets using an 80 percent and 20 percent split, respectively. These datasets are then used to develop and evaluate multiple predictive models, including k nearest neighbours regression, linear regression, decision tree regression, and support vector regression. In addition, long short-term memory networks are employed due to their suitability for modelling time dependent financial data. Model performance is assessed using standard error metrics, namely mean squared error, root mean squared error, and mean absolute error. These indicators evaluate the ability of the models to accurately predict stock index movements based on sentiment and textual information. A comparative analysis is conducted to examine the contribution of evolutionary optimisation in enhancing both machine learning and deep learning models, while minimising overfitting. Finally, the proposed approach is evaluated against existing forecasting methodologies applied during the pandemic period. By integrating evolutionary computation with machine learning and deep learning techniques, the hybrid framework provides a comprehensive understanding of the interaction between social media sentiment and stock market dynamics under conditions of heightened uncertainty such as those induced by the COVID-19 pandemic. The overall workflow of the proposed hybrid model architecture is illustrated in Figure 1.

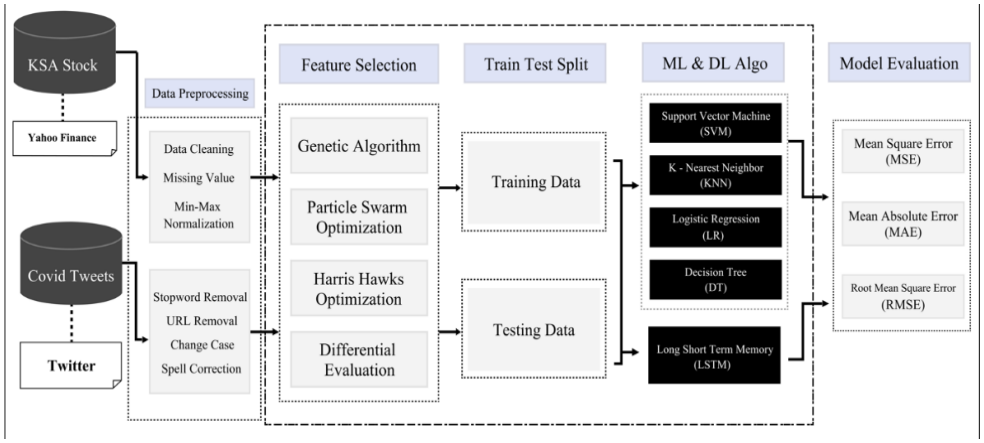

Fig.1: Proposed Model Diagram

3.1 Data Collection

In the contemporary digital environment, stock market behaviour can no longer be interpreted solely through conventional financial indicators. The widespread adoption of social media platforms has introduced an additional and influential dimension to market dynamics. This study adopts a comparative perspective to examine how the integration of traditional financial analysis with sentiment signals derived from social media can support more informed and effective decision making, thereby enhancing stock market forecasting and interpretative insight. By utilising financial data obtained from the Saudi Stock Exchange alongside sentiment information extracted from Twitter, the study seeks to capture latent market signals embedded in public opinion. The combined analysis enables the identification of meaningful sentiment driven patterns that have the potential to strengthen predictive models and improve the understanding of stock market movements.

3.1.1 Financial Data from the Saudi Stock Exchange (Yahoo Finance)

In this study, the primary dataset consists of historical financial records obtained from the Saudi stock exchange via the Yahoo Finance application programming interface. The data span an extended period from 2 March 2020 to 3 June 2023, allowing for a detailed examination of stock market trends and behavioural patterns over time. The data set contains essential pricing

information for each trading session, including opening, high, low, and closing prices. This comprehensive set of price indicators supports robust temporal analysis of market movements during the specified timeframe. A sample representative of the stock market dataset is presented in Table 1. This dataset serves as a foundational resource for assessing the performance and behavioural characteristics of Saudi equities during one of the most critical periods in recent history. It facilitates detailed investigation of stock price fluctuations, market volatility, and the potential influence of external factors, including economic shocks and geopolitical developments, thereby supporting a deeper understanding of the drivers shaping market dynamics.

Table 1
Examples of the Saudi Stock Dataset from Yahoo Finance

Date	Open	High	Low	Close	Adj Close
02/03/2020	3.7512	3.7512	3.746381	3.7512	3.7512
03/03/2020	3.746984	3.752	3.7454	3.747184	3.747184
04/03/2020	3.74802	3.7551	3.746224	3.74802	3.74802
05/03/2020	3.747896	3.7532	3.7444	3.747896	3.747896
06/03/2020	3.748653	3.7531	3.746954	3.748653	3.748653

3.1.2 Sentiment Data from Twitter

In this study, the second dataset comprises Twitter posts related to the Saudi stock market. These data were collected using a predefined set of keywords and hashtags associated with the COVID-19 pandemic and financial market activity, including "#COVID19," "#StockMarket," "#MarketCrash," "#StayAtHome," "#Investing," "#PandemicStocks," and "#VaccineNews." The selected hashtags were employed to extract tweets from the Twitter platform that reflect public discussions, opinions, and sentiment relevant to the Saudi stock market during the specified timeframe. A representative sample of the tweet dataset is presented in Table 2. The collected tweets are subsequently processed using a Sentiment Intensity Analyzer, which assesses the emotional tone expressed in each post concerning the Saudi stock market. Through this analysis, tweets are categorised into positive, negative, or neutral sentiments, enabling a clearer understanding of public perceptions and responses to events such as the COVID-19 pandemic, market fluctuations, investment strategies, and vaccine related developments. Integrating this sentiment annotated dataset with the financial data provides valuable insights into the influence of public opinion on the performance and behaviour of the Saudi stock market over the studied period. This approach demonstrates the utility of sentiment analysis in capturing how collective sentiments can shape market dynamics, inform investor behaviour, and affect overall market trends. By combining quantitative financial indicators with qualitative sentiment information, the analysis offers a comprehensive perspective on the interactions between external events, social media discourse, and market movements within the Saudi stock market context.

Table 2
Examples of Tweets Dataset in the Period of COVID-19

Date	Tweet
02/03/2020	The United fan boys cover their eyes in horror
03/03/2020	When everyone you know has died because of a treaty you was not involved in drawing up, you get an enormous amount of fury
04/03/2020	Sick of this shit. #mad #angry. Rowan Atkinson Is Not Dead
05/03/2020	Woke up feeling fresh with a clear mind.
06/03/2020	13th time seeing you guys today and you cancel the meet and greet because of the Covid

3.1.3 Integration and Analysis

This study explores the relationship between social media sentiment and stock market performance by integrating financial data from the Saudi Stock Exchange with sentiment information extracted from Twitter. Employing a combination of analytical techniques, the analysis aims to derive actionable insights from both structured financial data and unstructured social media content. Incorporating sentiment scores as additional features enhances the predictive accuracy and reliability of stock market forecasting models, allowing machine learning algorithms to account for the influence of public sentiment on market movements. This integrative approach facilitates the development of more robust predictive models that consider not only conventional financial indicators but also the nuanced impact of social media sentiment. The pre-processing steps applied to the Twitter dataset include:

3.2 Data Pre-Processing

1. Removal of Special Character and Punctuation: All special characters, symbols, and punctuation marks are eliminated from the tweets, as these elements introduce noise and contribute minimally to sentiment analysis.
2. Tokenization: Each tweet is segmented into individual words or tokens, enabling the assessment of sentiment at the word level.
3. Stop Word Removal: Common stop words, such as "a," "an," "the," "is," and "are," which carry little sentiment information, are removed to focus the analysis on meaningful terms.
4. Lowercasing: All words are converted to lowercase to ensure consistency in word representation, preventing identical words from being treated differently due to capitalization.
5. Lemmatization or Stemming: Words are lemmatized or stemmed to reduce them to their root form, decreasing the number of variants representing the same term and ensuring that related words are treated uniformly; for example, "running" and "ran" are both reduced to "run."
6. Emoticon and Emoji Handling: Emoticons and emojis are converted into textual representations to enable sentiment analysis algorithms to interpret and process them accurately.
7. URLs, Links, and User Mention Handling: URLs and user mentions are removed or replaced with generic tokens, as they generally do not convey sentiment and are considered noise.

Pre-processing steps for the stock exchange data:

Missing Value Handling: Identify variables with missing values in the dataset and address them using forward filling, backward filling, or interpolation techniques.

Data Integrity Check: Verify that the dataset is accurate and complete, adhering to the expected format and value ranges. This process involves detecting outliers, inconsistencies, or data entry errors and resolving them as necessary.

Data Scaling/Normalization: Ensure that numerical features are brought to a comparable scale or normalised where required. This prevents variables with large ranges from disproportionately influencing the analysis. Common approaches include min-max scaling or z-score normalization.

Feature Selection or Extraction: Identify and retain features that provide meaningful information for the analysis. This step reduces dimensionality and focuses on variables that significantly contribute to stock market evaluation.

Handling of Dates and Time: For datasets containing timestamps, extract relevant temporal features such as the day of the week or month, which can aid in detecting temporal patterns and trends in stock market data.

Irrelevant Column Removal: Remove columns that are unrelated to the analysis or are duplicated, thereby reducing data volume and improving computational efficiency.

These pre-processing steps for both the Twitter dataset and the stock exchange data establish a clean and well-prepared foundation for subsequent analysis. They ensure that sentiment analysis is accurate and facilitate meaningful insights into the relationship between social media sentiment and stock market dynamics.

3.3 Feature Extraction

Subsequently, the dataset undergoes processing using the TF-IDF technique for textual feature extraction. TF-IDF is a widely adopted method for evaluating the importance of terms within a set of documents, referred to as a corpus. The calculation of TF-IDF is expressed mathematically in Formula 1.

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t) \quad (1)$$

Here, TF-IDF (t, d) represents the TF-IDF score of term t in document d, where TF(t, d) denotes the frequency of term t within document d, and IDF(t) represents the inverse document frequency of term t. TF-IDF transforms the dataset by assigning higher weights to words that appear frequently in a particular document but are uncommon across the entire corpus. This weighting enables the identification of distinctive and meaningful terms that convey sentimental information pertinent to the sentiment analysis task. The TF-IDF method is critical in quantifying the significance of words within the dataset, facilitating the subsequent analysis of sentiment trends and patterns. By combining term prevalence with rarity across documents, TF-IDF provides a robust approach for extracting relevant textual features that enhance the accuracy and interpretability of sentiment analysis within the dataset.

3.4 Feature Selection

Feature selection is a critical step in any machine learning workflow, as it identifies the most relevant features from high-dimensional datasets, thereby enhancing model performance and interpretability. This study undertakes a rigorous comparative analysis of four swarm intelligence-based optimization algorithms—Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Harris Hawks Optimization (HHO), and Differential Evolution (DE)—for solving feature selection challenges in machine learning. These algorithms were chosen due to their established capacity to balance exploration and exploitation effectively across diverse problem spaces. Their robustness and adaptability have been demonstrated in various engineering and optimization contexts, particularly in addressing nonlinear, multimodal, and complex search spaces. Although newer metaheuristics such as Red Fox Optimization (RFO), Cuckoo Optimization (CO), and Reptile Search Algorithm (RSA) exist, they remain relatively untested compared with GA, PSO, HHO, and DE, whose effectiveness in analogous domains has been widely documented, providing a reliable basis for comparative evaluation.

The selection of these algorithms is further supported by Wolpert's No Free Lunch Theorem, which asserts that no single algorithm outperforms all others across every problem type, reinforcing the rationale for employing multiple metaheuristic approaches. Each algorithm draws inspiration from distinct natural processes, offering unique mechanisms for selecting relevant features while discarding irrelevant ones. To ensure robustness and statistical significance, all algorithms were executed thirty independent times, thereby accounting for stochastic variability and enabling a reliable assessment of performance across different runs. This study examines the specific strengths and limitations of each feature selection method, considering their suitability relative to dataset

characteristics and the machine learning task. The analysis also addresses implications for model interpretability and potential trade-offs between predictive accuracy and simplicity.

Through extensive experimentation across diverse datasets, the effectiveness, efficiency, and performance of these algorithms based on the selected feature subsets were systematically evaluated. The findings provide actionable insights into the advantages and disadvantages of each optimization algorithm, offering guidance for researchers and practitioners in selecting appropriate feature selection strategies to improve model accuracy and interpretability. By comparing and analysing multiple methods, this study establishes practical guidelines for choosing the most suitable approach for specific applications, thereby promoting the development of robust machine learning models and underlining the pivotal role of feature selection in enhancing both model performance and interpretability.

3.4.1 Genetic Algorithm (GA)

GAs are optimization techniques inspired by the principles of natural selection and genetics. They emulate the evolutionary process to identify optimal solutions by exploring a defined problem space. GAs operate on a population of candidate solutions, iteratively applying genetic operator selection, crossover, and mutation—to generate successive generations. This process enables the progressive refinement of solutions towards optimality. The fundamental workflow of a genetic algorithm is summarised in Figure 2.

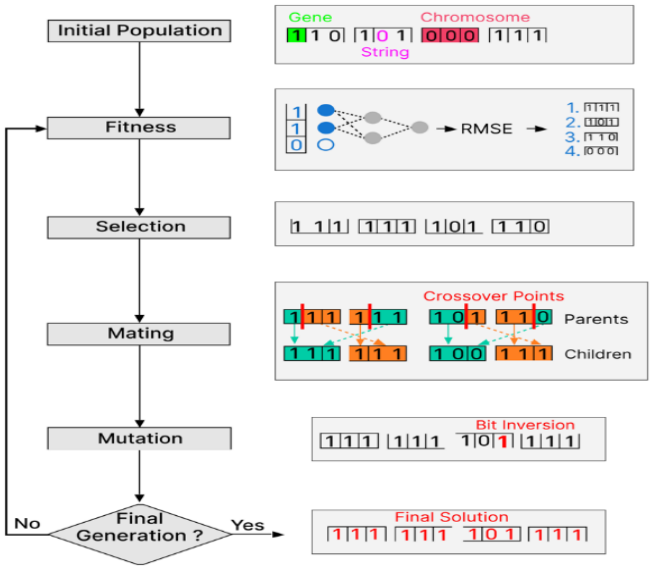

Fig.2: Genetic Algorithm Evolutionary Mutation Process

GAs is widely employed to address complex optimization problems across diverse domains, including machine learning, engineering, and economics. The core principle of GAs is to emulate natural evolution to identify the fittest solution for a given problem [34]. The process begins with an initial population of candidate solutions, generated randomly and represented as chromosomes, each corresponding to a potential solution within the search space. A fitness function evaluates each candidate by quantifying its performance relative to the optimisation objective. Selection then chooses individuals based on fitness, with higher fitness increasing the likelihood of selection, reflecting the principle of "survival of the fittest." The crossover operator combines genetic information from two selected candidates to produce offspring, introducing diversity into the population. Mutation further introduces random alterations to the genetic material of offspring,

limiting excessive convergence while maintaining variability. The new population comprises a combination of the best individuals from the previous generation and the newly created offspring, maintaining a constant population size.

This iterative process of selection, crossover, mutation, and replacement continues until convergence is achieved. Over successive generations, the population evolved towards increasingly optimal solutions, with the best candidates providing the optimal or near-optimal solutions for the problem at hand. In machine learning, GAs is applied for tasks such as feature selection, hyperparameter tuning, and model optimization. For feature selection specifically, GAs identifies the most informative subset of features that maximises model performance while reducing the total number of features, thereby improving both efficiency and predictive accuracy.

3.4.2 Genetic Algorithm Pseudocode

1. Initialization: Create the initial population of individuals: P = [P_1, P_2, ., P_n]
2. Fitness Evaluation: Evaluate the fitness of every individual in the population by F = [f(P_1), f(P_2), ., f(P_n)]
3. Selection: Choose individuals of the population for reproduction according to their fitness value: Selected = Select (P, F)
4. Crossover: Apply mutation to the offspring individuals to inject random changes: Mutated_Offspring = Mutate (Offspring)
5. Mutation: Apply mutation to the offspring individuals to introduce random changes: Mutated_Offspring = Mutate (Offspring)
6. Replacement: Replace some individuals in the population with the mutated offspring: P’ = Replace (P, Mutated_Offspring)
7. Termination: The iterative process from steps 2 to 6 continues until a predefined termination condition is satisfied, which may include reaching a maximum number of generations, achieving a satisfactory fitness level, or exceeding a specified time limit.

Here, P denotes the population, F represents the fitness values of the individuals, Selected refers to the individuals chosen for reproduction, Offspring indicates the new candidates generated via crossover, Mutated_Offspring represents the offspring after mutation has been applied, and P' denotes the updated population following the replacement process.

3.4.3 Particle Swarm Optimisation

This section outlines the PSO algorithm, focusing on its application to feature selection. The discussion includes the concept of particles, their positions, and velocities, as well as their interactions within the swarm to guide the search for optimal solutions. The overall procedure of the PSO algorithm for feature selection is summarised in Figure 3.

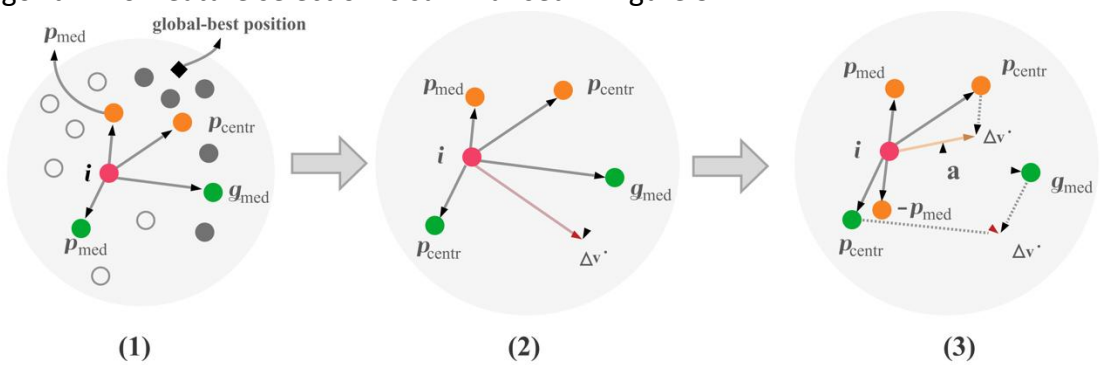

Fig.3: Particle Swarm Optimisation Evolutionary Process

PSO is an intelligent metaheuristic algorithm inspired by the collective behaviour observed in bird flocks or fish schools. It is widely applied to tackle complex optimization problems. The fundamental concept of PSO is to simulate the cooperative movement of a swarm of particles as they search for optimal solutions [35]. Each particle represents a potential solution within the search space and is defined by its position and velocity. Initially, particles are randomly distributed throughout the search space, with velocities assigned to dictate both direction and speed of movement. The optimization proceeds iteratively over a predetermined number of generations or until a convergence criterion is satisfied. In each iteration, the fitness of every particle is evaluated using a fitness function, which quantifies the quality of the solution represented by that particle, typically in terms of minimization or maximization of a target objective. During the search, particles adjust their positions and velocities based on two sources of information: their own best historical position (cognitive component) and the best position discovered by any particle in the swarm (social component). This collaborative exchange of information allows the swarm to converge efficiently towards superior solutions. The cognitive component directs each particle toward its individual's best position, while the social component guides the particle toward the global best position identified within the swarm. The following pseudocode provides a structured representation of the PSO algorithm for feature selection and other optimization tasks.

1. Initialization: Initialize the population of particles with their positions and velocities: $X = [X_1, X_2, ..., X_n]$ $V = [V_1, V_2, ..., V_n]$
2. Fitness Evaluation: Evaluate the fitness of each particle's current position: $F = [f(X_1), f(X_2), ..., f(X_n)]$
3. Update Personal Best (pBest): For each particle, update its personal best position if the current fitness is better than the previous personal best: pBest[i] = argmin(F[i], pBest[i])
4. Update Global Best (gBest): Find the particle with the best fitness among all particles: gBest = argmin(F)
5. Update Velocity and Position: Update the velocity and position of each particle based on its current velocity, position, pBest, and gBest: $V[i] = w * V[i] + c_1 * rand_1 * (pBest[i] - X[i]) + c_2 * rand_2 * (gBest - X[i])$ $X[i] = X[i] + V[i]$
6. Termination: Repeat steps 2-5 until a termination condition is met (e.g., maximum number of iterations or a satisfactory fitness level).

X represents the positions of particles, V represents the velocities of particles, F represents the fitness values of particles, pBest represents the personal best positions of particles, and gBest represents the global best position found by the swarm. The parameters w, c_1, and c_2 represent the inertia weight, cognitive parameter, and social parameter, respectively. $rand_1$ and $rand_2$ are random values between 0 and 1 used to introduce randomness in the velocity update equation.

3.4.4 Harris Hawks Optimization (HHO)

HHO, inspired by the cooperative hunting strategies of Harris hawks, is presented in this section. The discussion focuses on feature selection through the lens of hierarchical interactions, prey-capturing strategies, and information-sharing mechanisms inherent in the optimization process. The fundamental workflow of the HHO algorithm is summarised in Figure 4.

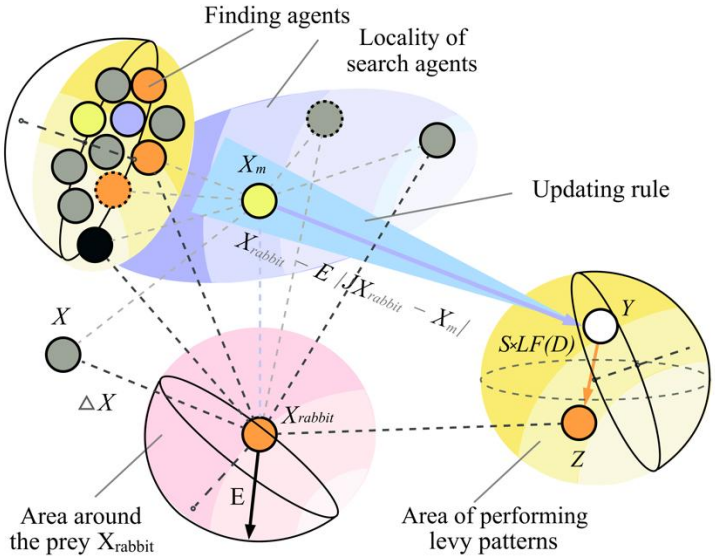

Fig.4: Harris Hawks Optimisation Algorithm Evolutionary Process

HHO draws inspiration from the social hierarchy and cooperative hunting strategies of Harris hawks in nature [36]. These birds exhibit coordinated and strategic behaviour during hunts, which forms the foundation of the HHO algorithm. Like other metaheuristic techniques, HHO begins with a population of hawks, each representing a potential solution within the search space, with its quality evaluated via a problem-specific fitness function. The positions of the hawks are iteratively updated over a set number of generations or until a predefined termination criterion is achieved. Each iteration of HHO involves two primary phases: exploration and exploitation. During exploration, certain hawks randomly traverse the search space to discover new promising regions, while others focus on exploitation, intensifying the search around areas deemed promising based on collective information from the swarm. Hawks in HHO are categorised according to their roles: explorers perform random searches, searchers concentrate on refining solutions in high-potential areas, and leaders guide the overall search process.

The position of each hawk is updated according to its role, its personal best position, the best position within its subgroup, and the global best position discovered by any hawk in the swarm. These updates are governed by mathematical models that integrate local and global information to guide the movement of the hawks. HHO is particularly effective at balancing exploration and exploitation, enabling efficient navigation of multimodal and boundedly rational search spaces. The hierarchical leadership structure facilitates information sharing and coordination among hawks, promoting faster convergence and comprehensive exploration of the search space. Over recent years, HHO has demonstrated strong performance across a range of optimization problems, including continuous function optimization, engineering design, and machine learning parameter tuning, establishing it as a robust member of the metaheuristic algorithm family. The pseudo-code for the HHO algorithm is presented below.

1. Initialization: Start by initializing the population of hawks' positions with $X = [X_1, X_2, ., X_n]$
2. Fitness Evaluation: The fitness of each hawk's position should be evaluated. Fitness evaluation is performed on each hawk's position by using the function f()
3. Update Leader Position: The hawk with the best fitness is identified as the leader: Leader = arg max(F)

4. Exploration Phase: Each hawk (other than the leader) The mean position of all other hawks except for the current hawk is given by: Mean = (ΣX - X[i]) / (N - 1), where N is total number of hawks.
5. Update the Location of the Current Hawk: X[i] = Mean + $rand_1$ * (X[i] - X[Leader]), where $rand_1$ is a random variable between 0 and 1.
6. Exploitation Phase: The position of the leader hawks is updated using the formula: X[Leader] = X[Leader] + $rand_2$ * (X[Leader] – Mean). The variable '$rand_2$' takes values between 0 and 1 randomly.
7. Termination: Repeat steps 2 through 5 until reaching a termination condition, for example, when reaching the maximal number of iterations or a satisfactory level of fitness.

In this context, X denotes the current position of a hawk, F represents the corresponding fitness value, Leader indicates the position of the leading hawk, and Mean corresponds to the average position of all hawks excluding the individual under consideration. The parameters $rand_1$ and $rand_2$ are employed to introduce stochasticity into the exploration and exploitation processes. Both $rand_1$ and $rand_2$ are random variables taking values within the interval [0, 1], thereby ensuring variability in the hawks' movements and preventing premature convergence.

3.4.5 Differential Evolution (DE)

DE is a population-based metaheuristic optimisation technique frequently applied to feature selection tasks [37]. The algorithm operates through three main mechanisms: the generation of mutant vectors, crossover between vectors, and a selection process to determine the fittest candidates. These steps collectively facilitate the identification of optimal feature subsets from high-dimensional datasets, enhancing the performance of machine learning models. The pseudo-code for implementing the DE algorithm is presented below.

Initialization: Initialize a population of candidate solutions: X = [X_1, X_2, ., X_n]

Mutation: For every candidate solution X[i]:

Select three different individuals randomly from the population: V = {V_1, V_2, V_3}, where $V_1 \neq V_2 \neq V_3 \neq X[i]$

Create a mutant vector by adding up the selected individuals using the formula: U = V_1 + F * (V_2 - V_3), where F is the scaling factor ($0 < F \leq 2$)

Recombination: For each candidate solution X[I]:

Create a trial vector by mixing the mutant vector U and current candidate solution X[i]: Y = X[i] + CR * (U - X[i]), with CR denoting the crossover rate (0 <= CR <= 1)

Selection: For each candidate solution Xi:

Assess the fitness of both X[i] and Y: F(X[i]) and F(Y)

Select the better solution based on fitness; That is, if F(Y) <= F(X[i]), then X[i+1] = Y; otherwise, X[i+1] = X

Termination: Repeat steps 2–4 until a predefined stopping condition is satisfied, such as reaching a maximum number of generations or achieving a satisfactory fitness level.

In this context, X represents the population of candidate solutions, V indicates the individuals selected for mutation, U corresponds to the generated mutant vector, Y denotes the trial vector, F is the scaling factor, and CR represents the crossover rate.

3.5 Ensemble Machine Learning Algorithms

Ensemble machine learning algorithms provide an effective approach for performing sentiment analysis and forecasting time-series related to Twitter discussions on the COVID-19 pandemic and stock-market activities. By combining the predictive capabilities of multiple base models, ensemble methods enhance the accuracy of sentiment classification and improve the reliability of stock-market trend predictions. This collective power offers valuable insights into both public sentiment and the underlying dynamics of financial markets. In the present study, the selected machine learning models include SVM, KNN, LR, and DT, chosen for their balance between accuracy, simplicity, and interpretability. While models such as XGBoost and AdaBoost might achieve higher predictive performance, their increased complexity is not necessary for the objectives of this research. Additionally, the chosen models have demonstrated strong performance in tasks aligned with the goals of the study.

3.5.1 SVM Regressor

The SVM Regressor is applied in this study to predict stock-market movements by leveraging sentiment analysis of tweets. The core objective is to utilise the SVM Regressor to identify and quantify the relationship between public sentiment expressed in tweets and the corresponding fluctuations in the stock market [38]. The accuracy of predictions generated by the SVM Regressor may vary due to the inherently volatile and dynamic nature of financial markets. It is also essential to emphasise that the selection of input data and the quality of pre-processing significantly influence the outcomes of the analysis. Furthermore, the predictive performance can be enhanced by incorporating additional explanatory variables, such as financial news or macroeconomic indicators, to provide a more comprehensive understanding of market dynamics.

3.5.2 KNN Regressor

The KNN Regressor is employed in this study to predict stock-market movements by utilising sentiment analysis from tweets. As a non-parametric machine learning model, KNN Regressor predicts outcomes by identifying the K nearest training instances to a given input and computing a weighted average of their target values [39]. The selection of an appropriate value for K is critical for model performance. A larger K value tends to produce smoother prediction curves but may fail to capture local trends, whereas a smaller K value may increase sensitivity to local variations yet introduce higher noise and potential bias from individual training instances. Consequently, it is necessary to experiment with multiple K values to determine the optimal configuration for the dataset under consideration. In applying KNN Regressor to stock-market prediction, it is important to note underlying assumptions, including stationarity of the dataset, reliability of sentiment analysis, and the influence of external factors on market trends.

3.5.3 LR Regressor

LR is applied to predict stock-market movements through sentiment analysis of tweets [40]. As a supervised learning algorithm, LR aims to model a linear relationship between independent variables and the target variables to forecast stock-market trends. The method assumes that such a linear association exists; however, stock-market behaviour is often nonlinear. Therefore, it is essential to recognise the limitations of LR in this context and consider alternative machine learning algorithms capable of handling nonlinear patterns for improved predictive performance.

3.5.4 DT Regressor

DT Regressor is employed to forecast stock-market behaviour using sentiment analysis derived from tweets [41]. As a widely used supervised learning algorithm, DTs can handle both categorical and continuous data features. A key advantage of DT Regressor is its interpretability, as its graphical structure clearly illustrates the decision-making process. However, DT Regressor is prone to overfitting, particularly when trees become deep and adapt to noisy data. Strategies to mitigate overfitting include pruning and employing ensemble techniques such as Random Forests.

3.6 Deep Learning (DL) Models

DL models are employed to predict the impact of news dissemination during the COVID-19 period on the stock market. As a specialised branch of machine learning, DL operates under the principles of artificial intelligence and mimics aspects of human brain functioning. In this study, the focus is on evaluating the effectiveness of DL models in processing large-scale datasets and identifying intricate patterns in news sentiment to forecast stock-market fluctuations. A detailed analysis of various DL architectures applied to predict stock-market movements in response to COVID-19 news demonstrates the models' capability in capturing complex sentiment-driven patterns, highlighting their effectiveness for such forecasting tasks.

3.6.1 Long Short-term Memory (LSTM)

In this study, LSTM neural networks were employed to forecast stock market trends using Twitter-derived data. LSTM, a specialised type of RNN, is particularly adept at handling sequential data and capturing long-term dependencies, enabling it to identify complex temporal patterns. The implementation involved training an LSTM network on a substantial dataset comprising historical stock market records alongside temporally aligned tweets relevant to those stock movements [42]. The input to the LSTM network consisted of features extracted from the tweets, reflecting key sentiment and market-related characteristics. These features were fed into the LSTM cells, allowing the network to process sequential information while retaining knowledge of preceding inputs. The training employed a supervised learning approach, incorporating historical stock market data to guide the network in learning associations between tweet-derived signals and market fluctuations. Once trained, the LSTM model was applied to unseen Twitter data to predict potential stock market movements, effectively capturing the influence of social media sentiment on market behaviour. The integration of Twitter-derived real-time information through LSTM networks provides a mechanism for investors and traders to access actionable insights, leveraging collective social sentiment to inform stock market decisions. This approach demonstrates the value of combining sequential modelling with social media analytics to enhance understanding of market dynamics in real time.

3.7 Performance Evaluation Measures

In this study, the importance of performance evaluation metrics is recognised as central to assessing the efficiency and accuracy of the proposed model. The application of such metrics enables a systematic and quantitative evaluation of the model's predictive capabilities, ensuring that its performance can be rigorously measured and validated.

3.7.1 Root Mean Square Error (RMSE)

RMSE represents the square root of the MSE and provides an estimate of the prediction error in the same units as the target variable. Lower RMSE values indicate better model performance, reflecting closer alignment between predicted and actual values. RMSE can be computed using Equation 2.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_{1,i} - x_{2,i})^2} \quad (2)$$

3.7.2 Mean Square Error (MSE)

MSE is a fundamental metric in regression analysis for assessing the accuracy of regression models. It quantifies the average squared difference between predicted values and the corresponding actual values. Lower MSE values indicate more precise predictions and better model performance. The MSE can be calculated using Equation 3.

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |x_{1,i} - x_{2,i}|} \quad (3)$$

3.7.3 Mean Absolute Error (MAE)

MAE measures the average absolute difference between predicted values and actual values, providing an indication of the typical magnitude of prediction errors without accounting for their direction. MAE is especially valuable when outliers might disproportionately affect the MSE. The MAE is calculated as shown in Equation 4.

$$MAE = \frac{1}{K}\sum_{j=1}^{k} |s_1 - s| \quad (4)$$

Where k denotes the number of errors and |s1-s| are the absolute errors.

4. Case Study

The COVID-19 pandemic has exerted substantial volatility on global financial markets, including the Saudi stock market, resulting in pronounced fluctuations in stock prices since early 2020. Such conditions have posed significant challenges for investors attempting to make informed decisions amidst the uncertainty generated by the pandemic's economic consequences. To address this real-world issue, this study focuses on predicting trends in the Saudi stock market during the COVID-19 pandemic by leveraging sentiment analysis of Twitter data. The primary objective is to employ social media sentiment as a valuable tool for gauging market mood, thereby enabling investors to make more informed decisions during these pandemic-induced uncertain periods. The methodology follows several critical steps. Initially, a large dataset of tweets related to the Saudi stock market was collected covering the period from 2 March 2020 to 3 June 2023. Tweets were retrieved using targeted hashtags and keywords such as "#COVID19," "#StockMarket," and "#VaccineNews," capturing public sentiments and discussions pertinent to the Saudi stock market within this timeframe. Subsequently, the collected tweets were processed using the VADER sentiment intensity analyser, which classified them as positive, negative, or neutral with respect to the Saudi stock market. This analysis provides insight into investor and public perceptions, reflecting the concerns, expectations, and sentiments prevailing during the pandemic.

Following sentiment classification, the Twitter data were integrated with historical financial data obtained from the Saudi stock exchange via the Yahoo Finance API. Exploratory analyses were conducted to identify correlations and patterns between tweet sentiments and stock-market

movements. Based on these insights, predictive models were developed using machine learning techniques, aiming to forecast stock-market trends informed by Twitter sentiment. These models were trained on historical market data combined with sentiment-labelled tweets, enabling them to anticipate future market movements under conditions like those induced by the pandemic. Finally, the predictive models were rigorously evaluated to assess their accuracy in forecasting stock market behaviour during the pandemic. The findings offer valuable guidance for investors, financial analysts, and policymakers, providing a means to make informed decisions amidst the uncertainties caused by COVID-19. These models also offer the potential for continuous monitoring of market trends, allowing stakeholders to track evolving sentiments and anticipate possible market shifts during ongoing or future economic crises. By applying sentiment analysis to social media data, this approach contributes substantively to the development of a resilient and adaptive financial market capable of navigating unprecedented global events.

5. Experimentation and Results

This section presents the experimental results obtained using the proposed model. In this study, ML, DL, and evolutionary algorithms were integrated to investigate the impact of Twitter content on stock indices. The dataset comprised stock-related sentiment tweets alongside stock exchange information spanning 2020–2021, obtained from Yahoo Finance. Initially, the data underwent pre-processing at multiple levels, including conversion to lowercase, tokenization, lemmatization, removal of stop words and links, and stemming. Feature extraction was subsequently performed using TF-IDF, resulting in a total of 59,589 features. Feature selection was then conducted by employing metaheuristic algorithms, specifically PSO, GA, DE, and HHO, to identify the most informative attributes. Following feature selection, various ML models were implemented, including KNNR, LR, DTR, and SVMR, alongside an LSTM model configured with ReLU activation and Adam optimisation, trained for 25 epochs. The dataset was divided into training and testing sets with an 80/20 split. Model performance was evaluated using MSE, RMSE, and MAE. The results demonstrate the effectiveness of evolutionary algorithms in enhancing feature selection, highlighting their ability to improve predictive performance. Furthermore, the findings provide valuable insights into forecasting stock-market trends during the COVID-19 period, emphasising the combined utility of sentiment analysis, ML/DL models, and metaheuristic feature selection in capturing complex market dynamics.

5.1 Data Visualisation

In this study, the tweets dataset was obtained from reliable sources, including Twitter and other verified news platforms. Correspondingly, the stock data utilised in the analysis were retrieved from Yahoo Finance. To maintain the relevance and alignment of the datasets, all data were synchronised by date, ensuring that each news article corresponded to the stock data of the same day. The data collection spanned from 2nd March 2020 to 3rd June 2023, providing a comprehensive timeframe for analysis. A visual representation of the dataset through a word cloud is presented in Figure 5.

Fig.5: Overview of Word Cloud for the Proposed Work

The NLP process produces the tokenised words illustrated in the word cloud presented in Figure 5. This visualisation highlights the most frequently occurring terms within the dataset, thereby emphasising their prominence and significance. Subsequently, the collected data were pre-processed following a series of steps incorporating NLP techniques. This preparation phase is of critical importance, as it ensures that the results of the analysis accurately reflect the underlying datasets and provide a robust foundation for reliable and valid research outcomes. Pre-processing primarily involves visualising the stock data using selected NLP techniques. The dataset encompasses daily stock information, including opening and closing prices, as well as the highest and lowest prices recorded each day. Initially, descriptive statistical measures are calculated for all variables, offering insights into the key characteristics of the stock data. These statistics are summarised in Figure 6, facilitating a comprehensive understanding of the dataset prior to further analysis and visualisation.

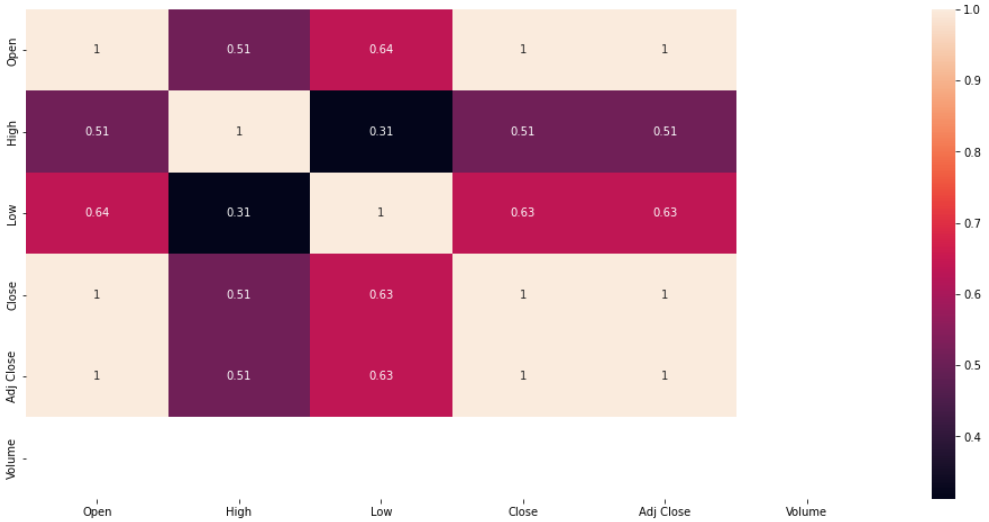

Fig6: Statistical Overview of Stock Data Characteristics

5.2 Experimental Results of Evolutionary Computation

This study introduces a hybrid methodology that integrates ML and DL techniques with evolutionary algorithms to examine the influence of Twitter content on stock indices. The dataset

comprises textual data from Twitter alongside stock exchange information retrieved from Yahoo Finance. The pre-processing procedures include lowercasing, tokenisation, lemmatisation, removal of stop words, elimination of links, and stemming, addressing potential issues in data quality. Feature extraction is conducted using the TF-IDF model, producing 59,589 features from the pre-processed dataset.

Subsequently, evolutionary algorithms—PSO, GA, HHO, and DE—are employed to identify the most relevant features, guided by iterations, generations, and fitness scores. The PSO algorithm selects 18,860 features, GA identifies 50,897 features, DE selects 59,723 features, and HHO chooses 51,453 features. These optimally selected features are then provided as input to ML models, including KNNR, LR, DTR, SVM Regressor, and the DL-based LSTM model. The dataset is partitioned into an 80% training set and a 20% testing set to facilitate model evaluation. The evolutionary algorithms are configured as follows: PSO with 30 particles and 30 iterations, GA with 30 chromosomes and 30 generations, DE with 30 particles and 30 iterations, and HHO with 30 chromosomes and 30 generations. The LSTM model is structured with two layers containing 120 and 190 neurons, utilising the ReLU activation function. Additionally, the LSTM component incorporates three dense layers with 50, 25, and 1 neuron, respectively. Training is conducted using the Adam optimiser, with MSE as the loss function, a batch size of 64, and 25 epochs. Model performance is assessed by using metrics such as MSE, RMSE, and MAE.

Table 3

Experimentation Results with Evolutionary Algorithms

Evolutionary Model	Type	Methods	Evaluation Metrics		
			MSE	RMSE	MAE
PSO	ML	KNNR	0.12810	0.00461	0.00249
		LR	1.79327	0.00461	0.45777
		DTR	2.75916	0.00525	0.00230
		SVR	2.72049	0.00521	0.00356
	DL	LSTM	0.00019	0.19937	0.17173
GA	ML	KNNR	0.41342	0.00272	0.00141
		LR	4.43789	0.21066	0.83074
		DTR	6.99216	0.00083	0.00010
		SVR	0.00015	0.01231	0.01187
	DL	LSTM	0.00020	0.20344	0.17862
HHO	ML	KNNR	0.65395	0.00128	0.00056
		LR	0.97535	0.30313	0.51422
		DTR	0.79702	0.00031	0.65750
		SVR	0.00014	0.00214	0.00170
	DL	LSTM	0.00029	0.24415	0.16022
DE	ML	KNNR	2.12810	0.06468	0.04440
		LR	0.31663	0.08031	0.18679
	DL	DTR	1.74567	0.34256	0.00216
		SVR	0.00536	0.08215	0.09169
		LSTM	0.0024	0.23164	0.16584

The results, analysed and compared, demonstrate the effectiveness of evolutionary algorithms in feature selection, as well as the predictive capability of the ML and DL models for stock-market forecasting. The outcomes are summarised in Table 3, providing insights into stock-market predictions based on news content during the COVID-19 period.

6. Conclusion

In this study, we propose a novel model for predicting stock-market movements based on news content sourced from online platforms, covering the full duration of the COVID-19 pandemic. Stock data were obtained from Yahoo Finance, while tweets were collected from Twitter. The primary focus of the proposed approach is to reduce model complexity and computation time, thereby achieving efficient yet highly accurate predictions. Model performance is evaluated using MSE, RMSE, and MAE, providing insight into the reliability and precision of the forecasts. Feature selection is facilitated through Evolutionary Algorithms, which identify the most informative and optimal features for stock trend prediction. Furthermore, a hybrid model combining HHO with SVR, referred to as HHO-SVR, is introduced, demonstrating superior predictive performance compared to existing approaches. By integrating evolutionary algorithms with machine learning and deep learning techniques, the proposed model enhances both the effectiveness and reliability of stock-market forecasts derived from news data. This framework provides investors and researchers with actionable insights into the relationship between news sentiment and market dynamics. Future work may explore the inclusion of newer metaheuristic algorithms, such as the Red Fox Optimizer and Crayfish Optimization, to further refine feature selection and improve model performance. Extending the analysis beyond the COVID-19 period would allow assessment of model robustness under diverse market conditions. Incorporating additional data sources, including financial statements and global economic indicators, could enhance predictive accuracy. Moreover, investigating alternative deep learning architectures and hybrid models may reduce computational overhead while maintaining or improving prediction performance.

Acknowledgement

The authors extend their appreciation to the Arab Open University, Saudi Arabia for giving resources for this work.

References

[1] Abdullah, M., Sulong, Z., & Chowdhury, M. A. F. (2023). *Explainable Deep Learning Model for Stock Price Forecasting Using Textual Analysis*. https://doi.org/10.1016/j.eswa.2024.123740

[2] Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2025). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research*, *345*(2), 1335-1386. https://doi.org/10.1007/s10479-021-04420-6

[3] Al-Hashedi, A., Al-Fuhaidi, B., Mohsen, A. M., Ali, Y., Al-Kaf, H. A. G., Al-Sorori, W., & Maqtary, N. (2022). Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories. *Applied Computational Intelligence and Soft Computing*, *2022*. https://doi.org/10.1155/2022/6614730

[4] Albahli, S. (2022). Twitter sentiment analysis: An Arabic text mining approach based on COVID-19. *Frontiers in Public Health*, *10*, 966779. https://doi.org/10.3389/fpubh.2022.966779

[5] Alotaibi, M. N., & Alharbi, Z. H. (2022). Sentiment analysis to explore user perception of teleworking in Saudi Arabia. *International Journal of Advanced Computer Science and Applications*, *13*(5). http://doi.org/10.14569/IJACSA.2022.0130565

[6] Alqarni, A., & Rahman, A. (2023). Arabic tweets-based sentiment analysis to investigate the impact of COVID-19 in KSA: a deep learning approach. *Big Data and Cognitive Computing*, *7*(1), 16. https://doi.org/10.3390/bdcc7010016

[7] Aslam, N., Khan, I.U., AlKhales, T., AlMakki, R., AlNajim, S., Almarshad, S. and Saad, R., . (2022). Sentiment analysis on Education transformation during Covid-19 using Arabic tweets

in KSA. *International Journal of Emerging Multidisciplinaries: Computer Science & Artificial Intelligence*, *2*(1), 35-45. https://ojs.ijemd.com/index.php/ComputerScienceAI/article/view/137

[8] Bacanin, N., Zivkovic, M., Jovanovic, L., Ivanovic, M., & Rashid, T. A. (2022). Training a multilayer perception for modeling stock price index predictions using modified whale optimization algorithm. In *Computational Vision and Bio-Inspired Computing*. Springer. https://doi.org/10.1007/978-981-16-9573-5_31

[9] Balaha, H. M., Antar, E. R., Saafan, M. M., & El-Gendy, E. M. (2023). A comprehensive framework towards segmenting and classifying breast cancer patients using deep learning and Aquila optimizer. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7897-7917. https://doi.org/10.1007/s12652-023-04600-1

[10] Balaha, H. M., El-Gendy, E. M., & Saafan, M. M. (2022). A complete framework for accurate recognition and prognosis of COVID-19 patients based on deep transfer learning and feature classification approach. *Artificial Intelligence Review*, *55*(6), 5063-5108. https://doi.org/10.1007/s10462-021-10127-8

[11] Balaha, H. M., & Saafan, M. M. (2021). Automatic exam correction framework (AECF) for the MCQs, essays, and equations matching. *Ieee Access*, *9*, 32368-32389. https://doi.org/10.1109/ACCESS.2021.3060940

[12] Boru İpek, A. (2023). Stock price prediction using improved extreme learning machine methods during the Covid-19 pandemic and selection of appropriate prediction method. *Kybernetes*, *52*(10), 4081-4109. https://doi.org/10.1108/K-12-2021-1252

[13] Chaouachi, M., & Chaouachi, S. (2020). Current covid-19 impact on Saudi stock market: Evidence from an ARDL model. *International Journal of Accounting Finance Auditing Management and Economics*, *1*(1), 1-13. https://revue.ijafame.com/index.php/home/article/view/8

[14] Elsegai, H., Al-Mutawaly, H. S., & Almongy, H. M. (2025). Predicting the Trends of the Egyptian Stock Market Using Machine Learning and Deep Learning Methods. *Computational Journal of Mathematical and Statistical Sciences*, *4*(1), 186-221. https://journals.ekb.eg/article_396438_0.html

[15] Fahmy, H., El-Gendy, E. M., Mohamed, M. A., & Saafan, M. M. (2023). ECH3OA: an enhanced chimp-harris hawks optimization algorithm for copyright protection in color images using watermarking techniques. *Knowledge-Based Systems*, *269*. https://doi.org/10.1016/j.knosys.2023.110494

[16] González-Núñez, E., Trejo, L. A., & Kampouridis, M. (2025). Expanding a machine learning class towards its application to the stock-market forecast. *Applied Intelligence*, *55*(1). https://doi.org/10.1007/s10489-024-06018-4

[17] Gülmez, B. (2025). GA-Attention-Fuzzy-Stock-Net: An optimized neuro-fuzzy system for stock-market price prediction with genetic algorithm and attention mechanism. *Heliyon*, *11*(3). https://doi.org/10.1016/j.heliyon.2025.e42393

[18] Jalil, Z., Abbasi, A., Javed, A. R., Khan, M. B., Hasanat, M. H. A., Malik, K. M., & Saudagar, A. K. J. (2022). Covid-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.812735

[19] Jana, R. K., & Ghosh, I. (2025). A residual driven ensemble machine learning approach for forecasting natural gas prices: Analyses for pre-and during-COVID-19 phases. *Annals of Operations Research*, *345*(2), 757-778. https://doi.org/10.1007/s10479-021-04492-4

[20] Kamali, A. H., Iranmanesh, S. H., & Goodarzian, F. (2024). Portfolio optimization in the stock-

market under disruptions: Real case studies of COVID-19 pandemic and currency risk. *Engineering Applications of Artificial Intelligence*, *136*. https://doi.org/10.1016/j.engappai.2024.108973

[21] Mahyoob, M., Algaraady, J., Alrahiali, M., & Alblwi, A. (2022). Sentiment analysis of public tweets towards the emergence of SARS-CoV-2 Omicron variant: A social media analytics framework. *Engineering, Technology & Applied Science Research*, *12*(3), 8525-8531. https://doi.org/10.48084/etasr.4865

[22] Mati, S., Ismael, G. Y., Usman, A. G., Samour, A., Aliyu, N., Alsakarneh, R. A. I., & Abba, S. I. (2025). Gaussian random fuzzy and nature-inspired neural networks: a novel approach to Brent oil price prediction. *Neural Computing and Applications*, 1-19. https://doi.org/10.1007/s00521-025-11306-2

[23] Mizdrakovic, V., Kljajic, M., Zivkovic, M., Bacanin, N., Jovanovic, L., Deveci, M., & Pedrycz, W. (2024). Forecasting bitcoin: Decomposition aided long short-term memory based time series modelling and its explanation with shapley values. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2024.112026

[24] Raghunathan, D., & Krishnamoorthi, M. (2025). Optimizing stock predictions with Bi-directional LSTM and levy flight fuzzy social spider optimization (LFFSSO): LSTM model. *International Journal on Semantic Web and Information Systems*, *21*(1), 1-25. http://doi.org/10.4018/IJSWIS.367280

[25] Saafan, M. M., & El-Gendy, E. M. (2021). IWOSSA: An improved whale optimization salp swarm algorithm for solving optimization problems. *Expert Systems with Applications*, *176*. https://doi.org/10.1016/j.eswa.2021.114901

[26] Vaiyapuri, T., Jagannathan, S. K., Ahmed, M. A., Ramya, K. C., Joshi, G. P., Lee, S., & Lee, G. (2023). Sustainable Artificial Intelligence-Based Twitter Sentiment Analysis on COVID-19 Pandemic. *Sustainability*, *15*(8). https://doi.org/10.3390/su15086404

[27] Wang, Y., Wei, W., Liu, Z., Liu, J., Lv, Y., & Li, X. (2025). Interpretable machine learning framework for corporate financialization prediction: A SHAP-based analysis of high-dimensional data. *Mathematics*, *13*(15), 2526. https://doi.org/10.3390/math13152526

[28] Wei, D., Wang, Z., & Qiu, M. (2025). Multiple objectives escaping bird search optimization and its application in stock-market prediction based on transformer model. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-88883-8

[29] Yang, C., Abedin, M. Z., Zhang, H., Weng, F., & Hajek, P. (2025). An interpretable system for predicting the impact of COVID-19 government interventions on stock-market sectors. *Annals of Operations Research*, *347*(2), 1031-1058. https://doi.org/10.1007/s10479-023-05311-8

[30] Zammarchi, G., Mola, F., & Conversano, C. (2023). Using sentiment analysis to evaluate the impact of the COVID-19 outbreak on Italy's country reputation and stock-market performance. *Statistical Methods & Applications*. https://doi.org/10.1007/s10260-023-00690-5

[31] Zivkovic, M., Stoean, C., Petrovic, A., Bacanin, N., Strumberger, I., & Zivkovic, T. (2021). A novel method for covid-19 pandemic information fake news detection based on the arithmetic optimization algorithm. 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE. https://doi.org/10.1109/SYNASC54541.2021.00051