Integrating Social Media Data and Historical Stock Prices for Predictive Analysis: A Reinforcement Learning Approach

Mei Li[1], Ye Zhang[2*]

College of Economic and Management, Jiangsu College of Engineering and Technology, Nantong 226000, Jiangsu, China[1]
College of Economic and Management, University of Padova, Padova 35100, Veneto, Italy[2]

***Abstract*—The reliance on data collection for assessing individual behavior and actions has intensified, particularly with the proliferation of digital platforms. People often use the Internet to express their opinions and experiences about various products and services on social media and personal websites. Concurrently, the stock market, a key driver of commercial and industrial growth, has seen a surge in research focused on predicting market trends. The vast array of information on social media regarding public sentiment towards current events, coupled with the known impact of financial news on stock prices, has led to the application of data mining techniques for understanding market volatility. This research proposes a novel method that integrates social media data, encompassing public sentiment, news, and historical stock prices, to predict future stock trends. The approach involves two primary phases. The first phase develops a sentiment analysis (SA) model using three dilated convolution layers for feature extraction and classification. Addressing the challenge of unbalanced classification, a reinforcement learning (RL)-based strategy is employed, wherein an agent receives varied rewards for accurate classification, with a bias towards the minority class. Additionally, a unique clustering-based mutation operator within a differential equation (DE) framework is introduced to initiate the backpropagation (BP) process. The second phase incorporates an attention-based long short-term memory (LSTM) model, merging historical stock prices with sentiment data. An experimental analysis of the study dataset is conducted to determine optimal values for significant parameters, including the reward function.**

Keywords—Social media; stock market; sentiment analysis; unbalanced classification; reinforcement learning; differential equation; long short-term memory

I. Introduction

Before the World Wide Web, people's decisions were influenced by opinions from colleagues and friends. With the rise of the internet, there's been an increase in sharing opinions with strangers online. People now express their views on a variety of topics, including news and products, on social networking sites like Twitter [1]. SA is a research field focused on categorizing these opinions into positive, negative, or neutral sentiments, often involving opinion mining to analyze textual expressions. Social media platforms like Facebook and Instagram have become key for expressing opinions, offering a vast resource for SA to gain insights into public sentiment on diverse topics [2].

In recent years, a specialized market has developed, focusing specifically on detecting and analyzing sentiments in the financial sector. This niche is centered on identifying and assessing sentiments tied to financial transactions. The skill to predict stock prices accurately is highly valuable for researchers and investors, especially given the volatility and unpredictability of these prices [3]. Machine Learning (ML) techniques have shown potential in improving the accuracy and dependability of stock market forecasts, yet designing an effective stock prediction model is still a complex task. Stock market behavior is shaped by a mix of social mood and historical price trends, which significantly influence stock price movements and changes. Integrating these elements into predictive models is essential for obtaining meaningful insights and making well-informed investment choices.

The influence of daily news articles on stock market performance is substantial [4]. These articles, containing vital information about companies, budgets, and trends, significantly affect public sentiment and stock market strategies. By analyzing these articles, investors can obtain critical insights for better investment decisions. This paper aims to use news articles to forecast stock market trends, focusing on those that provide insights into specific industries and predict future stock price movements. The increasing availability of financial data offers an opportunity to improve the speed and accuracy of these predictions [5].

SA faces challenges due to data imbalance, marked by significant discrepancies between negative and positive instances [6]. To combat this, there are both algorithm-based and data-based approaches. Data-based techniques include under-sampling, over-sampling, or combining both to reduce the impact of class imbalance. For instance, SMOTE [7] creates new samples by interpolating between minority instances, while NearMiss [8] applies under-sampling through the nearest neighbor algorithm. While over-sampling may lead to overfitting, under-sampling could result in the loss of important information. On the algorithm side, strategies are developed to give more weight to the underrepresented class. These involve enhanced ensemble learning, adjusting decision thresholds, and implementing cost-sensitive learning methods [9]. In cost-sensitive learning, classification is seen as a cost-minimization problem, assigning greater penalties for misclassifying minority samples. Ensemble techniques combine the predictions of multiple classifiers, and threshold adjustments modify the decision threshold during the testing phase.

Deep learning techniques, including Deep Reinforcement Learning (DRL), offer effective solutions for imbalanced classification issues [10, 11]. DRL stands out for its ability to manage imbalanced data through its distinctive features. It employs a reward system that prioritizes the minority class by either penalizing misclassification more severely or rewarding correct identification more generously. This method proactively counteracts the tendency of traditional models to favor the majority class. The benefits of DRL extend beyond just equalizing class distribution. It improves the recognition of significant patterns, especially those pertaining to the minority class, by adeptly filtering out irrelevant data. DRL's capacity to identify key, often-missed features in the dataset plays a vital role in developing models that are both more accurate and efficient [12].

Neural network weight initialization significantly affects the training efficiency and accuracy in SM prediction. Usually, weights are randomly assigned in gradient-based training algorithms like backpropagation [13, 14]. However, the choice of initial weights is vital for effective convergence and overall training performance [15, 16]. To improve weight initialization, population-based training can be used, selecting the optimal model from a pool to initiate the neural network. This method helps overcome the issue of local optima common in traditional methods. Notably, evolutionary algorithms have shown to be as effective as stochastic gradient descent for neural network training [17, 18]. Differential Evolution (DE) [19], a population-based optimization algorithm, is effective for machine learning weight initialization [20]. It offers several benefits: it explores the solution space with diverse candidate solutions, avoiding local optima and leading to optimal weight configurations. DE's iterative update process, based on the differential between target and current solutions, results in faster convergence and better performance. Additionally, DE is robust against noisy data, ensuring stable initial weight configurations even amidst data uncertainties. Its flexibility allows for customization to specific needs, like setting weight boundaries or incorporating prior knowledge, making it adaptable for various learning tasks and enhancing its effectiveness in weight initialization.

In this study, a novel approach is introduced to analyze social media data, combining public sentiment, personal opinions, news trends, and historical stock data to predict future stock prices. The methodology consists of two main stages. In the first stage, a SA model is developed using three dilated convolution layers to extract feature vectors and perform classification tasks. To address the challenge of unbalanced classification, a unique strategy based on RL is proposed, treating the task as a series of decision-making processes. The agent receives rewards at each step based on accurate classification, with a smaller reward assigned to the majority class to address the class imbalance. In the second stage, an attention-based Long Short-Term Memory (LSTM) approach is employed. This involves combining historical stock prices with the sentiment analysis results from the previous stage to make informed predictions. The optimal values for key parameters, including the reward function, are determined through experimental studies conducted on the dataset. To assess the individual contributions of the RL component, ablation studies are conducted to confirm the independent and cumulative positive effects on the overall performance of the model. The key contributions of the proposed model can be distilled into the following points:

- In the SA model, an ensemble of dilated convolutions is employed to extract valuable insights from textual data. This enhances the accuracy and informed decision-making in classification tasks.
- To address the challenge of imbalanced classification in SA, an RL strategy is proposed. This innovative approach provides a fresh perspective on achieving a balance between different classes.
- The model incorporates a unique reward system that reinforces correct decisions and penalizes incorrect ones. By assigning higher rewards to the minority class, the issue of dataset imbalance is effectively tackled, encouraging the model to give due attention to underrepresented data. This strategic approach contributes to a more balanced and equitable classification process.
- An improved DE algorithm that utilizes clustering to initialize weights in the SA model effectively has been devised. This approach aids in identifying a promising region to initiate the BP algorithm within the model. By selecting the optimal or near-optimal solution from the best cluster as the initial solution in the mutation operator and employing a new updating strategy, candidate solutions are generated more efficiently.

The structure of the article is as follows: Section II provides a literature review on stock market prediction. In Section III, the proposed approach is delved into in more detail. The experimental results and analyses are presented in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

Opinion mining focuses on capturing individuals' unique perspectives and opinions and finds applications in various fields such as product reviews, surveillance, healthcare, and politics. Accurately predicting changes in stock prices is a crucial research area, and recent advancements have been made in creating predictive models for the global stock market. Traditional methods like time series analysis and machine learning (ML) models are commonly used in academic research to analyze stock market predictions. Numerous approaches for analyzing social media (SM) have been explored in scholarly literature. Milosevic et al. [21] and Pandya et al. [22] introduced an ML method that utilizes selected financial variables to predict long-term investments, achieving improved performance through feature selection. Porshnev et al. [23] combined support vector machine (SVM) and neural network (NN) algorithms, utilizing a lexicon-based approach to analyze psychological states and their impact on stock market indicators like DJIA and S&P500. They effectively demonstrated the influence of Twitter data on stock market performance. Lai et al. [15] investigated stock prediction models using support vector machine (SVM) and least square SVM. Athale et al. [24] and Mehta et al. [25] utilized SA and ML to explore the relationship between public sentiment and stock performance, aiming to address the complex and unpredictable nature of non-linear and non-parametric financial time series. Xing et al. [26] highlighted the capabilities of natural language processing (NLP) in financial forecasting, particularly in the domain of Natural Language-Based Financial Forecasting (NLFF) or sentiment analysis using data from SM platforms. NLP techniques are rapidly advancing, specifically in NLFF and the analysis of SM data, driven by practical applications in financial forecasting. However, despite these advancements, a gap remains in integrating comprehensive social media data with advanced sentiment analysis and machine learning techniques for precise stock market predictions. The proposed research aims to bridge this gap by combining public sentiment, news, and historical stock prices to predict future stock trends more accurately.

Emotion plays a crucial role in facilitating effective communication, as indicated by research conducted by Pandya et al. [22] and Smailović et al. [27], who explored the usefulness of Twitter feeds for predicting stock closing prices and assessing public sentiment towards companies and their products. In this paper, a method is introduced to measure the probability of positive and negative expressions or opinions, with the aim of forecasting SA in the field of finance. Using the Granger causality test, stock price trends can be examined over a short period, providing valuable insights from the data [5]. The use of SVM enables the categorization of tweets into positive, negative, or neutral sentiments, thereby enhancing the predictive capability of SM platforms. Various initiatives have been undertaken to predict SM trends, with a focus on accurately forecasting the significant impact of a company's SM presence through real-time collection of Twitter data. The study demonstrates the effectiveness of sentiment analysis in extracting public mood from Twitter and other SM platforms to predict fluctuations in individual stock prices. In order to bolster the analytical process, the active learning model has been integrated with a stream-based method. This combination enables the algorithm to choose fresh training data, thereby enhancing its performance. The financial implications of this analysis are evaluated by employing Recurrent Neural Networks (RNNs) in a trial-based methodology. The proposed model extends these efforts by employing a novel clustering-based mutation operator within a DE framework and integrating sentiment analysis with attention-based LSTM models, a distinct approach not yet explored in existing literature.

Examining sentiment helps in evaluating the impact of emotions within a textual context when analyzing decision-making with a positive outlook. Bhuriya et al. [28] developed predictive models using regression techniques to forecast the market price of TCS. These models were based on attributes such as large price, close price, open price, small price, and volume. The study assessed the effectiveness of linear, polynomial, and radial base function regression models by considering the confidence values associated with the predicted results. Despite having expertise, predicting stock prices remains challenging due to unexpected fluctuations influenced by historical trends and past price changes in the global market economy. Barot et al. [4] proposed a research approach that utilized sentiment analysis to analyze news articles and forecast stock price fluctuations. They introduced a method for categorizing articles based on sentiment, classifying them as positive, negative, or neutral. This approach was integrated into machine learning models to improve the accuracy of stock market prediction. Patel et al. [29] and Patel et al. [30] examined the practical implementation of a machine learning model to forecast stock and share price movements in the Indian equity market. The analysis incorporated four forecasting models (SVM, ANN, Naïve Bayes, and Random Forest) with two input methods. Khedr et al. [31] introduced an optimized model that aimed to reduce error ratio and improve the accuracy of predicting patterns in share price performance. Their predictive model integrated sentiment analysis of financial news and historical values of social media data. These approaches have successfully utilized various types of market and company data, leading to superior outcomes compared to previous research. Chen et al. [32] conducted a comparison between conventional neural network price prediction and deep learning techniques using Chinese social media data on CSI 300 shared values. Their findings indicated that deep learning predictions outperformed conventional neural networks in terms of performance. Carosia et al. [33] conducted a study to investigate the influence of SM activities, specifically on the Brazilian SM platform Twitter, on the market value of specific companies. They utilized sentiment analysis (SA) to analyze the impact of SM movements on these companies. The study considered three different perspectives: the total count of emotions expressed in tweets, tweet sentiments weighted by the number of favorites, and tweet sentiments weighted by the number of retweets. SA was performed using the Multilayer Perceptron technique to analyze sentiment in Portuguese. Deep learning algorithms, including deeply convolutional neural networks (CNN), have gained prominence in SA. Santos et al. [34] employed a deep CNN trained on the Stanford Sentiment Treebank and the Stanford Twitter Sentiment Corpus. The model achieved an impressive accuracy of over 85% in predicting sentiment for both datasets. Yoshihara et al. [35] discussed an SA technique that combined recurrent neural networks (RNN) with Deep Belief Networks (DBN) in deep learning, resulting in improved financial market forecasting with reduced error rates compared to SVM and DBN strategies. In another study by Jiang et al. [36], deep learning models were analyzed to forecast SM trends. The research involved categorizing various NN models, evaluating metrics, exploring their integration, and assessing their reliability. Pang et al. [37] and Sun et al. [38] proposed the utilization of STM in NN techniques to enhance the accuracy of financial market predictions, particularly for the Shanghai A-shares composite index. Khan et al. [39] incorporated data algorithms that combined social networks and business news to evaluate the impact on the accuracy of SM forecasts over a 10-day period. They implemented feature selection and spam tweet removal techniques, and the use of deep learning techniques improved the accuracy rates. The research findings indicated that assessing the impact of SM is a complex task, with New York and Red Hat stocks showing significant sensitivity to SM activity, while London and Microsoft stocks are more influenced by financial news.

To sum up, existing literature showcases a variety of methods for price prediction. Although traditional deep learning techniques have significantly progressed the field, their limitations include a lack of social media data utilization and sensitivity to initial weights, which hinders their practical application in forecasting. Moreover, the issue of unbalanced classification remains a prevalent challenge for many deep learning models. Addressing these shortcomings, the study introduces an innovative methodology that combines the strengths of RL with the adaptive potential of a DE algorithm, specifically fine-tuned for initial weight optimization. This approach is engineered to surmount the conventional obstacles encountered by deep learning methods, thereby boosting the model's adaptability and efficiency in identifying cost-effective strategies for price prediction. The goal is to offer a sophisticated and robust tool to the suite of price prediction methods, one that is not only theoretically ground-breaking but also practically relevant in a variety of real-world scenarios.

III. Proposed Method

Based on Fig. 1, the proposed model is structured into four distinct steps, each integral to the overall process:

- The first step is Data Collection, where relevant data is gathered from various sources.
- The second step is Data Pre-processing. In this phase, the collected data undergoes cleaning and transformation to make it suitable for analysis.
- In the third stage, the focus shifts to analyzing news content using NLP for SA. This essential phase leverages advanced natural language processing (NLP) methods to assess the emotional undertones within the news material. The aim is to systematically identify the sentiment of each article, classifying it as either positive, negative, or neutral. The SA model is specially tailored to refine the classification mechanism, addressing critical challenges such as class imbalance and the need for accurate initial weight configurations. The integration of DE and RL in the proposed framework specifically targets these pivotal areas, which are often inadequately addressed by existing models. Conventional techniques typically lack a structured approach for setting initial weights, potentially slowing down the learning process and leading to suboptimal solutions. In the context of SA, where promptness and precision in prediction are essential, these limitations can be significant. Moreover, the model's RL component is designed to provide greater incentives for correctly predicting less frequent classes, thus shifting the focus to these vital predictions. This represents a marked advancement over traditional supervised learning models, which may struggle with insufficient data representation across all categories. The flexibility of RL in adjusting the learning strategy ensures a more equitable exploration of decision-making options, fostering approaches that prioritize the accurate identification of less represented categories. This distinctive capability of RL within the model distinguishes it from current methods, enabling it to address the unique challenges that traditional classification models face in SA applications.
- The final step is SM forecasting. Using the insights gained from the sentiment analysis, along with other relevant financial indicators and historical data, this stage focuses on predicting future movements of the stock market.

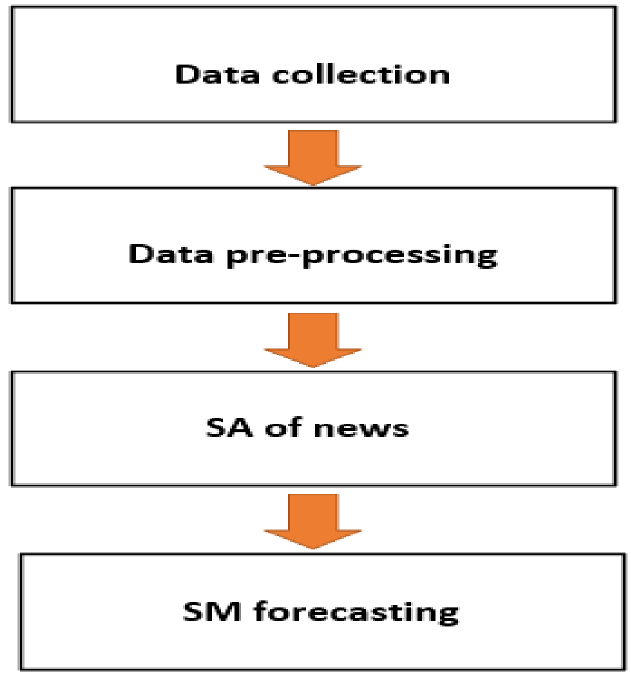


Fig. 1. The constituent stages of the proposed model.

A. Data Collection

The system design is founded on the use of two distinct datasets from various data providers. The first dataset includes news articles from reputable platforms like Moneycontrol, IIFL, Economic Times, and Twitter. This collection provides a broad perspective on market-related news, encompassing company announcements, industry trends, and economic developments, crucial for understanding stock market behavior. The second dataset contains historical records from the National Stock Exchange (NSE) of India, spanning six years. It features key parameters such as date, time, open, high, low, close values of stocks, and trading volume. This dataset is instrumental in analyzing past market trends and patterns, thereby aiding in generating insights into the behavior of individual stocks and the overall market. The data management strategy involves storing the gathered information in CSV file format. This approach offers several benefits. Firstly, it ensures convenient organization, enabling users to navigate and interpret the data efficiently. Secondly, the reduced file size makes it more manageable and less resource-intensive to handle large amounts of data. Additionally, the ease of generating and manipulating CSV files facilitates smooth data processing. Finally, this format enhances data integrity, ensuring that the information remains accurate and reliable for analysis. By integrating these datasets, the system aims to create a comprehensive framework that leverages the strengths of both news and stock market data. This dual-dataset approach is expected to provide a more nuanced and detailed understanding of the factors influencing the stock market, thus supporting more informed and effective decision-making in financial analytics.

B. Data Pre-processing

The pre-processing phase of the study begins with synchronizing text data and converting it to lowercase characters. This step is vital to address various text factors that could impact the classification process. The primary objective is to prepare the input data for sentiment classification, making it suitable for analysis. This preparation involves several tasks, such as removing links, special symbols, and emoticons, along with eliminating stop words. Further, the process includes analyzing parts of speech, applying stemming techniques, and tokenizing the text. These steps are essential to extract meaningful features for sentiment analysis. The output from the pre-processor subsystem then becomes the input for the sentiment classifier method.

The framework described in the article adopts a data-driven approach, leveraging historical stock prices recorded at opening and closing times. This rich dataset forms the foundation for analyzing and understanding stock price behavior over time. To represent these prices effectively, the framework utilizes a sentiment polarity approach. This involves categorizing stock prices as either positive or negative, aiming to capture the underlying emotional sentiment in the stock market. The sentiment polarity representation is instrumental in providing insights into market dynamics, highlighting the influence of emotions and perceptions on stock prices. To substantiate the effectiveness of this approach, a study was conducted at Stanford University [40]. This study integrated sentiment polarity values derived from the framework, aiming to explore the correlation between sentiment polarity and stock price movements. The findings of this study are enlightening, revealing the intricate relationship between market sentiment and financial outcomes. They underscore the importance of sentiment analysis in predicting stock market trends. These insights are particularly beneficial for investors, traders, and financial analysts, aiding them in making more informed and strategic decisions.

C. SA of News

Fig. 2 in the study illustrates the proposed structure for the SA model. This model functions based on a sentence D, defined as a series of words $[w_1, w_2, \ldots, w_\mathrm{n}]$, where n is the maximum number of words in a sentence. The sentence is fed into the BERT model, which generates an embedding matrix $E = [e_1; e_2; \ldots; e_n]$, with each e_i representing the embedding of the corresponding word w_i. To extract features, three dilated convolution layers are applied in parallel to matrix E. Each of these layers independently extracts a distinct feature vector from the sentence. Following this, max pooling is implemented to capture the most significant features while simultaneously reducing computational complexity. The outputs from the max-pooling layers are then directed into a Multilayer Perceptron (MLP) for classification purposes. The MLP's output is a vector of length three, representing the classification of the input sentence into one of three categories: positive, negative, or neutral. However, one of the challenges encountered in this model is the imbalanced nature of the dataset, predominantly skewed towards sentences classified as positive. This imbalance can potentially impair the performance of the system. To address this imbalance, the model incorporates an imbalanced classification Markov decision process. This process involves the use of a sequential decision-maker, specifically designed to tackle the challenges posed by the imbalanced classification problem. The decision-maker operates by adjusting the classification strategy, focusing on equitably distributing attention across all classes, including those less represented in the dataset. This approach not only enhances the accuracy of the model but also ensures a more balanced and equitable classification outcome, vital for the integrity and applicability of sentiment analysis in various contexts.

1) Pre-training: Weight initialization plays a vital role in the performance of deep learning models [41]. Choosing the right initial values for weights is crucial because inappropriate values can lead to convergence problems, making the model less effective or even rendering it non-functional. Proper initialization helps in achieving a faster convergence rate and improves the overall training efficiency of the model. It also impacts the ability of the model to reach global or good local minima in the optimization landscape. Different initialization methods, such as Xavier/Glorot, He initialization, or random initialization, are tailored for specific types of neural networks and activation functions. These methods aim to maintain a balance in the variance of the activations and gradients throughout the network, which is critical for deep models. The right choice of weight initialization method can significantly influence the learning dynamics and the eventual success of the model in tasks like classification, regression, or feature learning.

DE [42] stands as a prominent method within the realm of evolutionary algorithms, primarily used for solving complex optimization problems. As a population-based approach, DE focuses on iteratively refining a group of candidate solutions to converge towards an optimal solution for a given problem. The strength of DE lies in its unique mechanisms of mutation, crossover, and selection, which collectively drive the evolutionary process. Mutation in DE is a distinctive process where a new candidate solution is generated by adding the weighted difference between two randomly selected solutions from the population to a third solution. This approach introduces diversity and aids in exploring the solution space. The crossover step in DE further enhances solution diversity by combining features of the mutated solution with an existing member of the population, creating a trial solution [43]. This crossover mechanism ensures a mix of characteristics from different solutions, promoting the exploration and exploitation of the solution space. Selection, the final step in DE, is critical for the evolution of the population. It involves a comparison between the trial solution and the existing solution, retaining the one that offers better performance according to the defined objective function. This selective pressure gradually improves the overall quality of the population, steering it towards the optimal solution. DE's effectiveness is not limited to a single domain but extends across various fields such as engineering, economics, and machine learning. Its popularity stems from its simplicity, efficiency, and versatility in handling different types of optimization problems, including those with nonlinear, multimodal, and high-dimensional characteristics. Its ability to find high-quality solutions with relatively simple operations and fewer control parameters makes it a go-to choose for practitioners and researchers in optimization tasks.

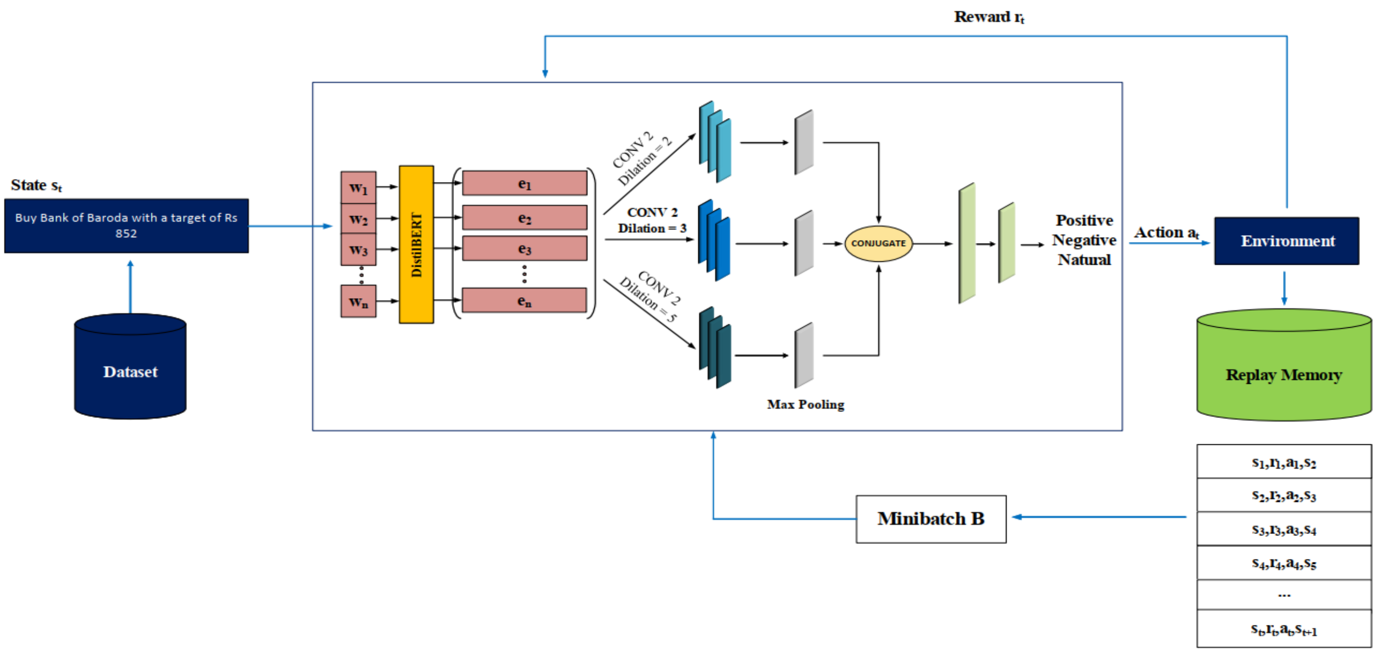


Fig. 2. The proposed SA model.

The mutation operator in DE is a pivotal component that significantly influences the algorithm's ability to search effectively across the solution space. This operator is responsible for introducing variability in the population, which is essential for exploring new regions in the search space and avoiding local optima. However, the effectiveness of the mutation operator is a delicate balance. If it is overly aggressive, it might disrupt the population's structure too much, leading to the loss of promising solutions and potential premature convergence or stagnation. In contrast, an overly conservative mutation approach may not provide sufficient exploration, leading to slow convergence rates and possibly settling for suboptimal solutions. To tackle these challenges, various modifications and enhancements to the standard mutation process in DE have been proposed. One of the notable advancements is the introduction of adaptive mutation schemes. These schemes dynamically adjust the mutation parameters, like the mutation rate and scaling factor, based on real-time feedback from the population's performance. By doing so, the algorithm can maintain an optimal balance between exploration and exploitation throughout the optimization process, adapting to the changing landscape of the search space. Another innovative approach is self-adaptive DE, where the mutation parameters are not fixed but are treated as part of the solution itself. In this method, each candidate solution carries its mutation rate and scaling factor, which evolve alongside the solution. This self-adaptation allows for a more tailored mutation process for each candidate solution, potentially enhancing the diversity and adaptability of the population [44].

To further boost the performance of the DE algorithm, a cutting-edge mutation and updating approach, which incorporates clustering concepts, is utilized to bolster the optimization process. This innovative strategy, drawing inspiration from the methodologies outlined in reference [45], focuses on employing the mutation operator to precisely target designated sectors within the search landscape. This is achieved through the application of the k-means algorithm, which effectively divides the current population P into k distinct clusters. Each cluster corresponds to a unique segment of the search space, thus facilitating a more focused and efficient exploration. The determination of the number of clusters is conducted in a stochastic manner, with the possible range stretching from 2 to the square root of N ($\sqrt{N}$). Upon completion of the clustering phase, the algorithm proceeds to identify the most optimal cluster. This is determined based on the average fitness level of the members within each cluster, with the one displaying the lowest mean fitness being designated as the most favorable. This particular cluster is then prioritized in the subsequent phases of the algorithm, as it is indicative of a region in the search space with a higher potential for yielding optimal solutions. Fig. 3 serves as a visual representation of this advanced methodology. In this figure, a theoretical problem encompassing 19 potential solutions is illustrated. These solutions are methodically grouped into three separate clusters, each representing a different area of the solution space. This visualization not only demonstrates the clustering mechanism but also highlights the effectiveness of this approach in partitioning the population for more targeted and efficient optimization. Through this sophisticated clustering-based mutation and update strategy, the DE algorithm is significantly enhanced, offering a more robust and precise tool for tackling complex optimization challenges.

The proposed mutation, based on clustering, is then defined as [45]:

$$\overrightarrow{v_i^{clu}} = \overrightarrow{win_g} + F(\overrightarrow{x_{r_1}} - \overrightarrow{x_{r_2}}) \tag{1}$$

In this context, $\overrightarrow{x_{r_1}}$ and $\overrightarrow{x_{r_2}}$ symbolize two randomly chosen candidate solutions from the existing population, whereas $\overrightarrow{win_g}$ represents the optimal solution found within the identified promising region. It is critical to acknowledge that $\overrightarrow{win_g}$ may not always equate to the best solution across the entire population.

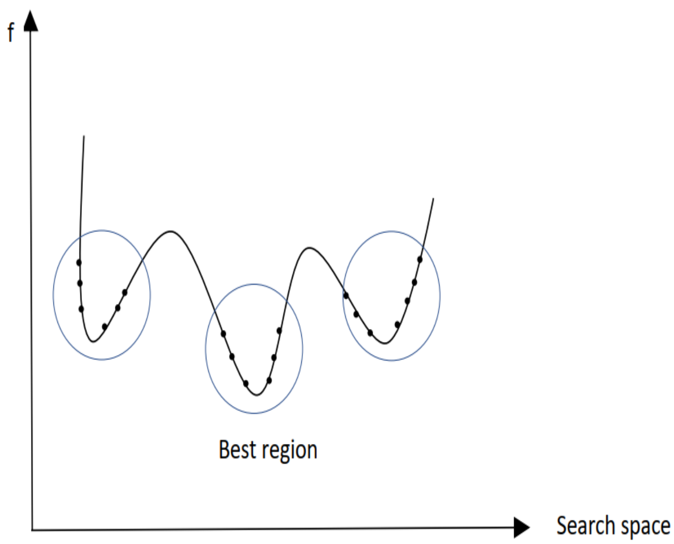


Fig. 3. Utilizing population clustering within the search space to identify the most favorable region.

Subsequent to the creation of M novel solutions via the mutation process influenced by clustering, the current population undergoes an update process as per the guidelines of Gradient-based Population Adjustment (GPBA) [46]. GPBA is a sophisticated mechanism designed to refine the population in evolutionary algorithms like DE. GPBA works by evaluating the gradient information of the current population. This involves assessing how each candidate solution is positioned relative to the objective function's gradient. By doing so, GPBA can determine the direction in which the solutions should be adjusted to move closer to the optimal point in the search space. This gradient-based approach differs from traditional evolutionary strategies, which rely solely on fitness-based selection and random mutations. The key advantage of integrating GPBA into the DE algorithm is the enhanced convergence speed towards optimal solutions. Since GPBA utilizes gradient information, it can guide the population more effectively towards the global optimum, especially in complex, multimodal landscapes where traditional methods might struggle. In the context of the DE algorithm, after the clustering-based mutation process identifies promising regions and generates new solutions, GPBA takes over to fine-tune these solutions. It adjusts the population by moving the candidate solutions along the gradient of the objective function. This not only accelerates the convergence process but also helps in maintaining diversity within the population, preventing premature convergence to local optima.

The procedure for this update is outlined as follows:

- Selection: Generate k random individuals to serve as the initial seeds for the algorithm.
- Generation: Produce a set of M solutions using mutation based on clustering and denote it as v^{clu}.
- Replacement: Choose M solutions randomly from the current population to form set B.
- Update: Choose the best M solutions from the union of sets v^{clu} and B to form set B'. The new population is obtained by combining the elements of set P that are not in B with the elements of a set B' $((P - B) \cup B')$

2) Deep q-network training: In sentiment analysis, unbalanced classification occurs when the distribution of sentiment labels in the dataset is highly skewed, with one sentiment class being more prevalent than the others. This poses challenges as standard classification algorithms tend to favor the majority class, leading to poorer performance for the minority class [47, 48]. Consequently, models might excel at predicting the majority sentiment but struggle with accurately identifying and classifying minority sentiments. The imbalanced nature of the dataset can introduce bias and impact the learning process, resulting in models with lower recall or sensitivity for the minority class. This is particularly problematic when the minority class represents important sentiments, such as negative feedback in customer analysis. To address the challenge of imbalanced classification, a sequential decision-making strategy utilizing RL is adopted. In RL, an agent engages with its environment, aiming to maximize cumulative rewards through the selection of optimal actions. Within the devised model, the agent acquires a sample from the dataset and undertakes a classification task at each sequential time step. Subsequently, the agent obtains immediate feedback from the environment: accurate classifications are rewarded with positive scores, while erroneous classifications incur negative scores. This approach ensures dynamic learning and adaptation by the agent, enhancing its decision-making capabilities in imbalanced classification scenarios [49, 50]. Suppose a dataset of N samples is available, each with corresponding labels $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)\}$, where x_i represents a sample and y_i denotes its label. The upcoming details provide the planned arrangements for the proposed approach.

- Policy π_θ : Policy π serves as a function which links states (S) to corresponding actions (A), with $\pi_\theta\ (s_t)$ denoting the action taken in a particular state s_t. The classification technique that employs the weights θ is identified as π_θ.

- State s_t : Each sample x_t from the dataset, D is associated with a specific state s_t. The beginning state s_1 is symbolized by the first piece of data x_1. To deter the model from acquiring a particular sequence, D is shuffled in every episode.

- Action a_t: The action a_t is performed to estimate the label x_t, where the categorization is binary, and a_t can take on either the value of 0 or 1. In this context, 0 signifies the less prevalent class, whereas 1 symbolizes the more common class.

- Reward r_t : The reward is determined by how well the action is executed. If the agent correctly classifies, it is given a positive reward; conversely, if it errs, it is penalized with a negative reward. The reward value should be varied for each class. By appropriately calibrating rewards, the model's performance can be significantly improved by ensuring that the reward magnitude matches the corresponding action. The method for determining the reward for a given action in this study is described by the following equation:

$$r_t(s_t, a_t, y_t) = \begin{cases} +1\,, a_t = y_t \; and \; s_t \in D_{Maj} \\ -1\,, a_t \neq y_t \; and \; s_t \in D_{Maj} \\ \lambda\,, a_t = y_t \; and \; s_t \in D_{Min} \\ -\lambda\,, a_t \neq y_t \; and \; s_t \in D_{Min} \end{cases} \quad (2)$$

Here, D_{Maj} and D_{Min} denote the majority and minority classes correspondingly. Accurately or inaccurately classifying a sample from the majority class results in a reward of $+\lambda$ or $-\lambda$, respectively, where λ is a value between 0 and 1.

- Terminal E: Throughout every training episode, the training process culminates at different terminal states. A series of state-action pairs $\{(s_1, a_1, y_1), (s_2, a_2, y_2), (s_3, a_3, y_3), \ldots, (s_t, a_t, y_t)\}$ from the starting state to the ending state is known as an episode. In the context of the situation, the conclusion of an episode is marked either by classifying all the training data or by misclassifying a sample from the minority class.

- Transition probability P: The agent advances from the present state, s_t, to the subsequent state, s_{t+1}, following the order of the data being read. The likelihood of this transition is denoted as $p(s_{t+1}|s_t, a_t)$.

D. SM Forecasting

LSTM networks have garnered considerable attention in stock market analysis due to their ability to capture and leverage temporal relationships in time series data [51]. Unlike traditional statistical models, LSTM networks excel at learning and adapting to complex patterns and dependencies in the data, making them well-suited for predicting stock market prices or trends. One advantage of LSTM networks is their proficiency in handling long-term dependencies, enabling them to capture subtle relationships and trends that span extended periods. By analyzing historical price data, trading volumes, and other relevant factors, LSTM networks can identify recurring patterns, seasonal trends, and market cycles that traditional models might miss [52]. Another strength of LSTM networks is their capacity to process and retain information over long sequences. The architecture of an LSTM cell includes memory units that selectively remember or forget information over time, enabling the network to store and retrieve relevant historical data while disregarding irrelevant or redundant information. This memory retention mechanism is crucial in capturing the dynamics of stock market movements, where past trends and events significantly impact future outcomes.

Additionally, LSTM networks can effectively handle irregularities and non-linearities in stock market data. The stock market is influenced by various factors, such as economic indicators, geopolitical events, news, and investor sentiment, leading to abrupt shifts and anomalies in stock prices. LSTM networks, with their ability to model complex non-linear relationships, can effectively capture and respond to these non-linear dynamics, aiding in market prediction.

The LSTM model, initially conceptualized by Hochreiter and Schmidhuber [53], has evolved significantly through various advancements since its inception [54]. In this particular study, the focus is on the highly regarded LSTM architecture developed by Gers et al. [55], a structure that has seen extensive application in diverse academic research [56, 57]. The fundamental aspect of the LSTM model is its distinctive gating mechanism, which effectively manages the storage and retrieval of information over temporal sequences. This research specifically explores the intricacies of the LSTM cell, in line with the delineation presented in Graves' work [36]. LSTM's architecture is characterized by a set of three specialized gates that orchestrate the efficient processing of information. These gates, denoted as i_t (input gate), f_t (forget gate), and o_t (output gate), function at a given time t in the sequence. They work in concert with c_t (memory cell) and h_t (hidden state), which are crucial components of the LSTM's internal structure. Furthermore, x_t symbolizes the external input received by the LSTM cell at the specific time t. These components collectively enhance the LSTM's capacity for learning, retaining, and applying long-term data dependencies, distinguishing it from conventional recurrent neural network designs. Its proficiency in grasping long-term dependencies renders the LSTM highly appropriate for intricate tasks associated with sequential data. This includes areas such as time series forecasting, natural language processing, and speech recognition, where context and historical data are crucial for precise modeling and prediction. The operational framework of the LSTM cell is structured in the following manner [56]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t tanh(c_t) \quad (7)$$

The sigmoidal activation function, represented by the symbol σ(.), plays a pivotal role in the computational models presented. This function, which includes both logistic sigmoid and hyperbolic tangent types, is applied in a manner that targets individual elements. The weight matrices, identified as W_{xk}, where k is a member of the set $\{i, f, o, c\}$, are integral to the process, linking the input x_t with distinct components such as the input gate, forget gate, output gate, and memory cells. In addition to these, the W_{ck} weight matrices, where k falls within the set $\{i, f, o\}$, are structured as diagonal matrices. These matrices are essential in forming the connections between the memory cell and its various gates. It is vital to highlight that the neuron count designated for each gate is set beforehand. This predefined count is significant because Eq. (3) to Eq. (7), which are central to the model, are applied distinctly to each neuron, ensuring precise and individualized processing within the neural network system.

IV. Empirical Evaluation

A. Metrics

In this article, three standard metrics are used for SM prediction: root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE) [58]. Each metric is defined in the following manner:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (8)$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (9)$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N} \quad (10)$$

Here, N stands for the total number of observations, with y_i representing the actual value and $\hat{y}_i$ the predicted value for each observation. Furthermore, for SA, Measures like Accuracy, F-measure, and G-means are employed, each defined in a specific manner [59]:

$$\text{Accuracy} = \frac{TP + TN}{Total\ number\ of\ samples} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

$$\text{G} - \text{means} = \sqrt{\text{Recall} \times Specificity} \quad (15)$$

Here, TP stands for true positives, which are the actual positive cases that the model has correctly identified. TN is for true negatives, representing the actual negative instances that the model has accurately predicted. FP refers to false positives, where the model mistakenly labels actual negatives as positive. Finally, FN denotes false negatives, pertaining to the actual positive cases that the model incorrectly categorizes as negative.

B. Comparator Models

The introduced algorithm was subjected to a thorough comparative analysis against eight state-of-the-art models, including those proposed by Zhang et al. [60], Rasheed et al. [61], Jin et al. [62], Vijh et al. [63], Mehta et al. [64], Riady [65], Wang et al. [66], and BL et al. [67]. This thorough analysis was designed to provide a comprehensive view of the effectiveness of the proposed model compared to current methods. Additionally, the algorithm is compared with a derivative model called "Proposed without RL," which excludes RL for SA. The objective of this study was to validate the superior performance and effectiveness of the proposed model in the domain of stock market prediction.

C. Results

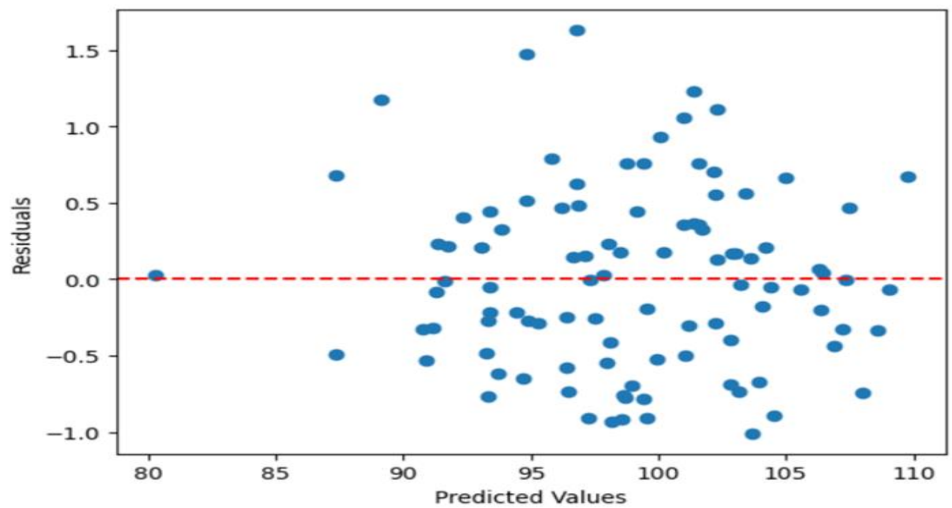
The comparative results, summarized in Table I, clearly demonstrate the exceptional performance of the proposed algorithm, consistently surpassing the other models evaluated on a shared dataset. Remarkably, the proposed algorithm outperformed the Transformer-based model by Wang et al., reducing overall error by approximately 12% and confirming the superior efficacy of attention-based LSTM over transformer. Furthermore, the model surpassed the model by BL et al., the strongest competitor, across all evaluation criteria. The performance improvement of the model is particularly significant when considering the substantial reduction in error rate. Specifically, the proposed algorithm achieved a remarkable decrease in error for two primary evaluation metrics, RMSE and MAPE, with errors reduced by over 14% and 10% respectively. This substantial reduction in error highlights the superior predictive capabilities of the algorithm. A separate comparison with the "Proposed with RL" model emphasized the critical role of RL in the proposed model. The comparison revealed an approximate 10% reduction in error rate when RL was implemented. These findings underscore the potency of RL and LSTM in navigating the complexities of time-series data, such as stock market prices, and their essential contribution to enhancing the accuracy of the proposed model.

Residual plot is a valuable graphical tool frequently employed in statistical and regression analyses. This type of plot visualizes the differences between observed and predicted values, known as residuals, on the vertical axis, while the predicted values are displayed on the horizontal axis. In Fig. 4, residual plots are presented for the models detailed in Table I.

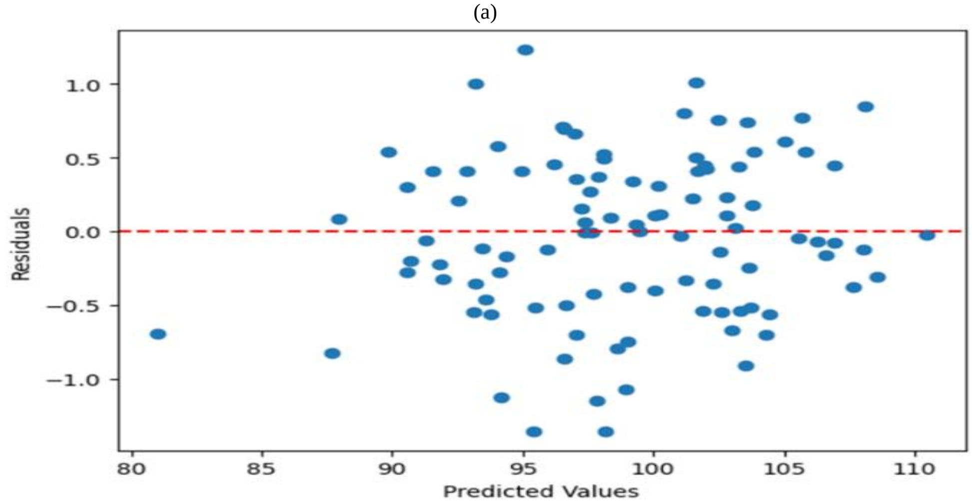
In the case of the proposed method, the data points predominantly cluster around the zero point, indicating minor discrepancies between the observed and predicted values. This clustering suggests a high level of accuracy in the model predictions. Additionally, the randomness and uniform distribution of the points demonstrate that the residuals are independent and identically distributed, a characteristic referred to as homoscedasticity. This important feature confirms that the linear regression model adheres to its assumptions, ensuring unbiased and reliable predictions. The absence of discernable patterns in the plot, such as curvilinear trends or funnel shapes, indicates that the model effectively captures the linear relationship between the predictors and the outcome variable. Moreover, this absence implies a consistent variance in the error term across various predicted value levels. These characteristics signal that the proposed model has successfully captured the underlying trends in the data and provides robust predictions.

TABLE I. RESULTS OBTAINED USING THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS

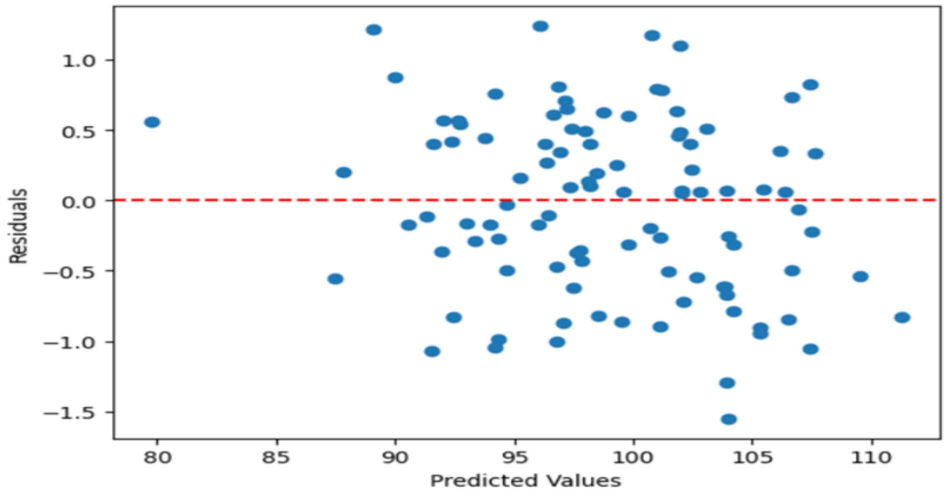
Model	RMSE	MAPE	MAE
Zhang et al. [60]	6.100	0.0410	4.100
Rasheed et al. [61]	5.402	0.0375	4.140
Jin et al. [62]	5.416	0.0302	3.715
Vijh et al. [63]	5.025	0.0260	3.493
Mehta et al. [64]	4.125	0.0241	3.201
Riady [65]	4.001	0.0217	4.128
Wang et al. [66]	3.852	0.0187	3.040
BL et al. [67]	3.501	0.0162	2.963
Proposed without RL	3.742	0.0137	2.825
Proposed	2.147	0.0125	2.130



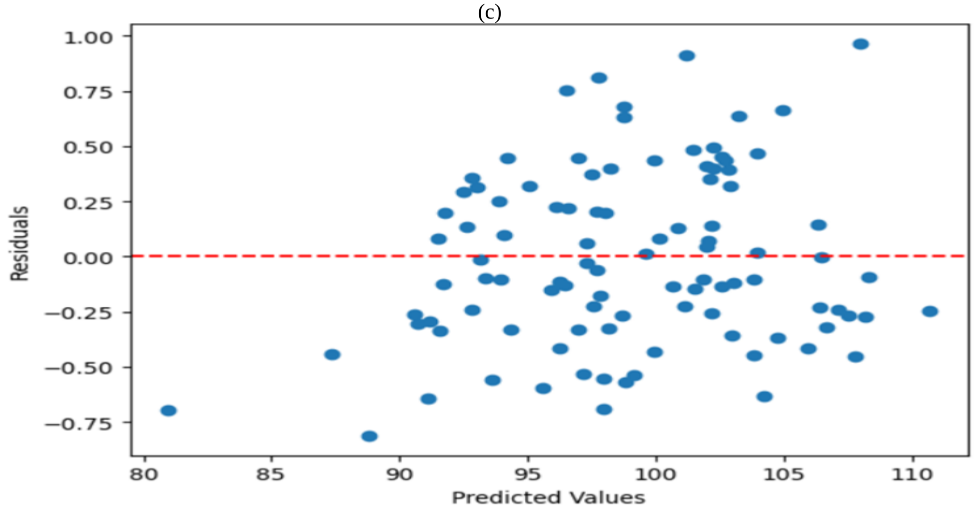
(a)



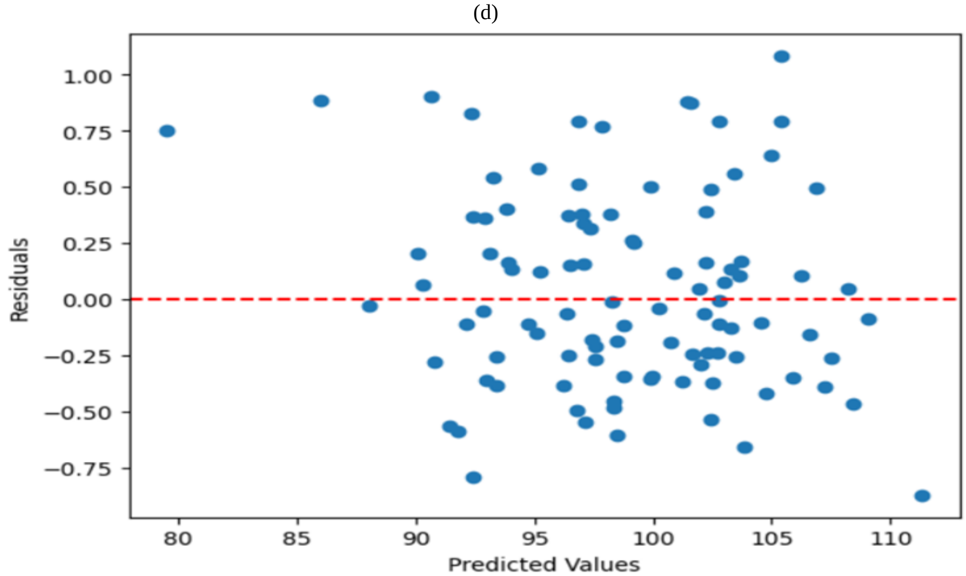
(b)


Residuals
Predicted Values

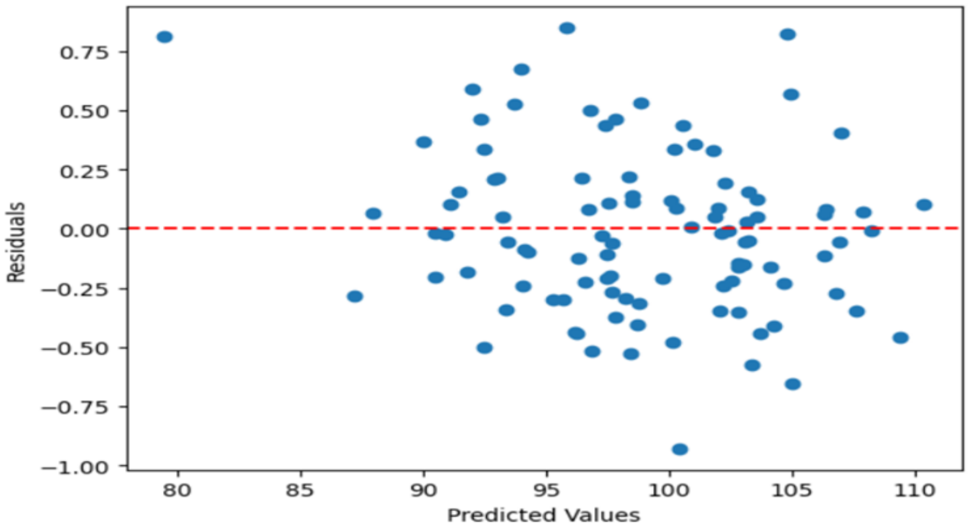
(c)


Residuals
Predicted Values

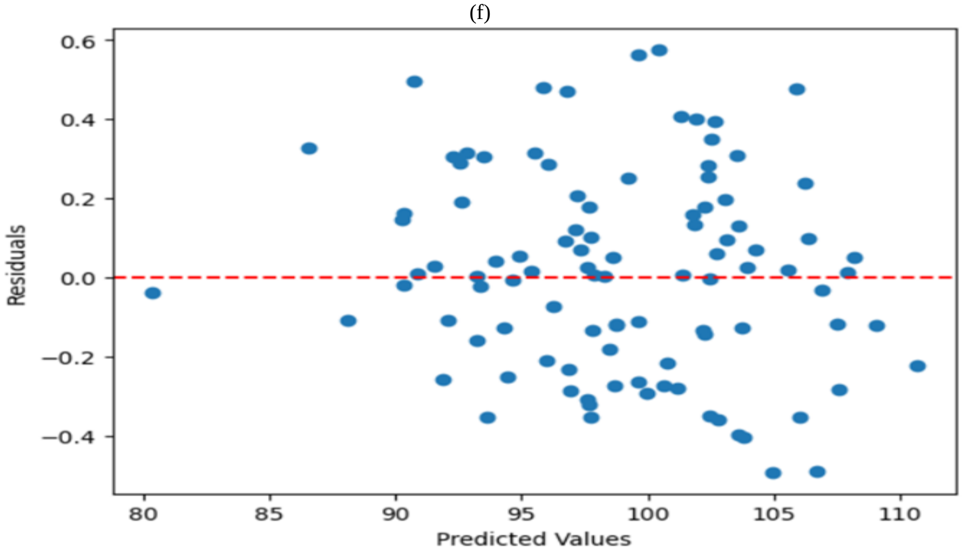
(d)


Residuals
Predicted Values

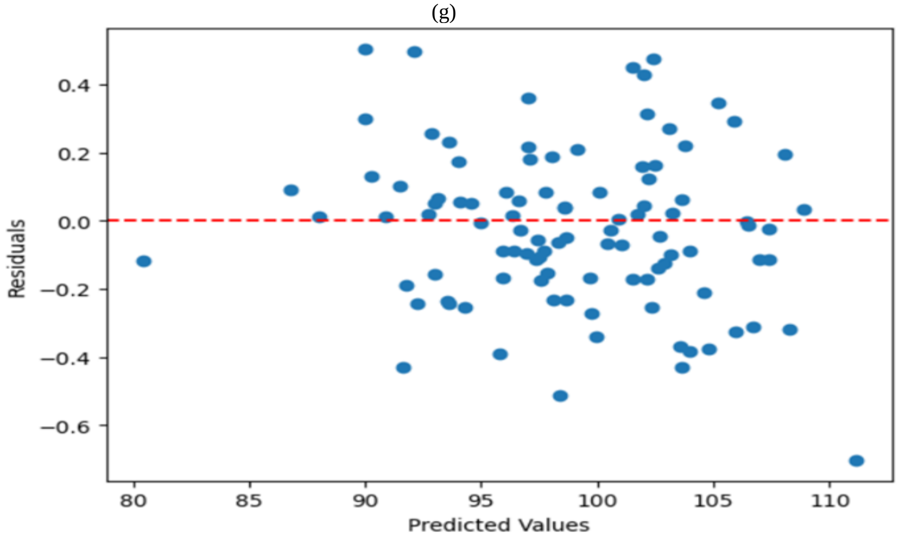
(e)


0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Residuals
80
85
90
95
100
105
110
Predicted Values

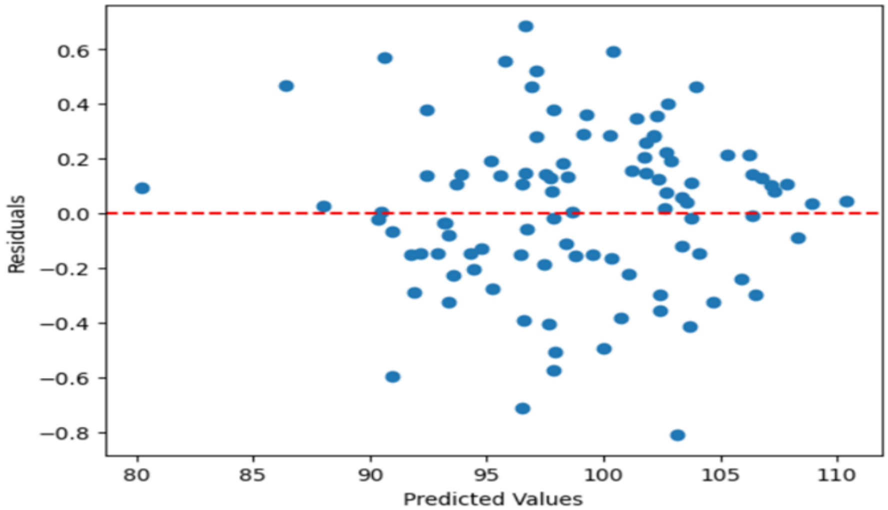
(f)


0.6
0.4
0.2
0.0
−0.2
−0.4
Residuals
80
85
90
95
100
105
110
Predicted Values

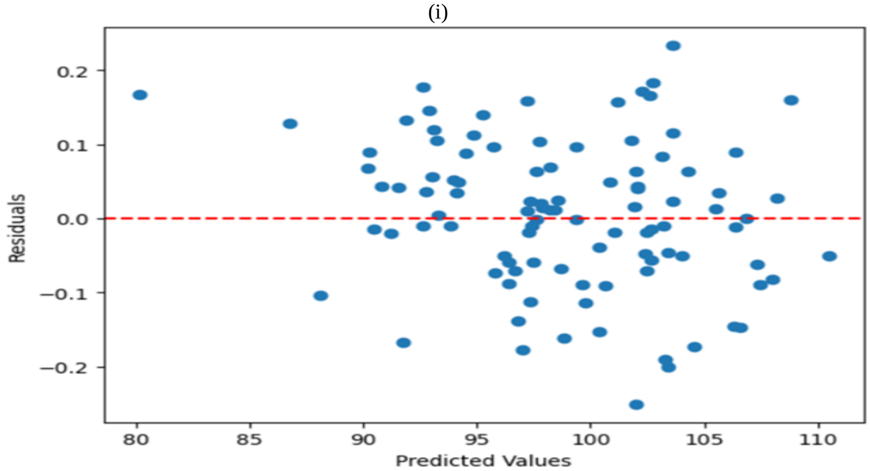
(g)


0.4
0.2
0.0
−0.2
−0.4
−0.6
Residuals
80
85
90
95
100
105
110
Predicted Values

(h)



(i)



(j)

Fig. 4. Residual plot for the proposed model and other state-of-the-art models. a) Zhang et al. [60], b) Rasheed et al. [61], c) Jin et al. [62], d) Vijh et al. [63], e) Mehta et al. [64], f) Riady [65], g) Wang et al. [66], h) BL et al. [67], i) Proposed without RL, j) Proposed.

1) Performance of semantic analysis: In the following analysis, the objective is to juxtapose the performance of the proposed SA model against five other prominent SA models. The comparative study includes models developed by Akhtar et al. [68], Akhtar et al. [69], Silva et al. [70], Xin et al. [71], and Mingzheng et al. [72], each of which holds considerable acclaim and widespread application in the SA domain (see Table 2). In assessing the results of the model, standard performance metrics such as F-measure and G-mean are employed, recognized for their dependability in evaluating imbalanced data [73]. Notably, the proposed model surpassed all other models across all evaluative parameters, even outpacing the highest-performing model, Mingzheng et al. More specifically, the proposed approach cut the error rate by an impressive 32% and 25% in the F-measure and G-mean metrics respectively. Further scrutiny was applied by comparing the performance of the proposed model with a condensed version of the proposed approach, referred to as proposed without RL. This comparison uncovered that the fully developed model drastically curtailed the error rate by an estimated 51%. These compelling findings underscore the critical importance and effectiveness of the RL technique embedded in the proposed approach.

Fig. 5, which illustrates the receiver operating characteristic (ROC) curves for the methodologies listed in Table II, employs the area under the curve (AUC) as a crucial metric for assessing classifier efficiency. An AUC of 1 is indicative of impeccable differentiation capabilities, whereas a score of 0.5 equates to a performance level similar to random guessing. Impressively, the method we developed showcased exceptional proficiency, achieving an AUC of 0.61. This score not only reflects its ability to accurately distinguish between positive and negative outcomes but also reinforces the method's validity as an effective predictive instrument. Furthermore, the 'Proposed without RL' variant demonstrated substantial efficacy, achieving an AUC of 0.53. This performance confirms its capacity to accurately classify positive and negative cases. In contrast, the methodologies developed by Mingzheng et al. and Xin et al. recorded more modest AUC values of 0.50, demonstrating a performance level that does not quite reach the benchmark set by our proposed approach. This ROC analysis distinctly highlights the varying levels of effectiveness among the different methodologies examined.

The notable predictive strength of our proposed method, both independently and in conjunction with RL, is a testament to its robustness and reliability in predictive modeling. This success not only validates the approach but also suggests the possibility for future enhancements. The method's adaptability and high level of accuracy pave the way for its potential application in a broader range of predictive scenarios, offering promising prospects for advancements in the field of predictive analytics. This creates exciting opportunities for further research and development, potentially leading to even more refined and efficient predictive models in the future.

TABLE II. RESULTS OBTAINED USING THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART MODELS FOR SA

Model	Accuracy	F-measure	G-means
Akhtar et al. [68]	0.695 ± 0.160	0.580 ± 0.041	0.695 ± 0.160
Akhtar et al. [69]	0.705 ± 0.015	0.580 ± 0.035	0.705 ± 0.015
Silva et al. [70]	0.825 ± 0.165	0.760 ± 0.263	0.825 ± 0.255
Xin et al. [71]	0.800 ± 0.015	0.700 ± 0.120	0.800 ± 0.000
Mingzheng et al. [72]	0.855 ± 0.105	0.820 ± 0.056	0.855 ± 0.269
Proposed without RL	0.860 ± 0.005	0.850 ± 0.012	0.860 ± 0.035
Proposed	0.891 ± 0.015	0.890 ± 0.003	0.898 ± 0.055

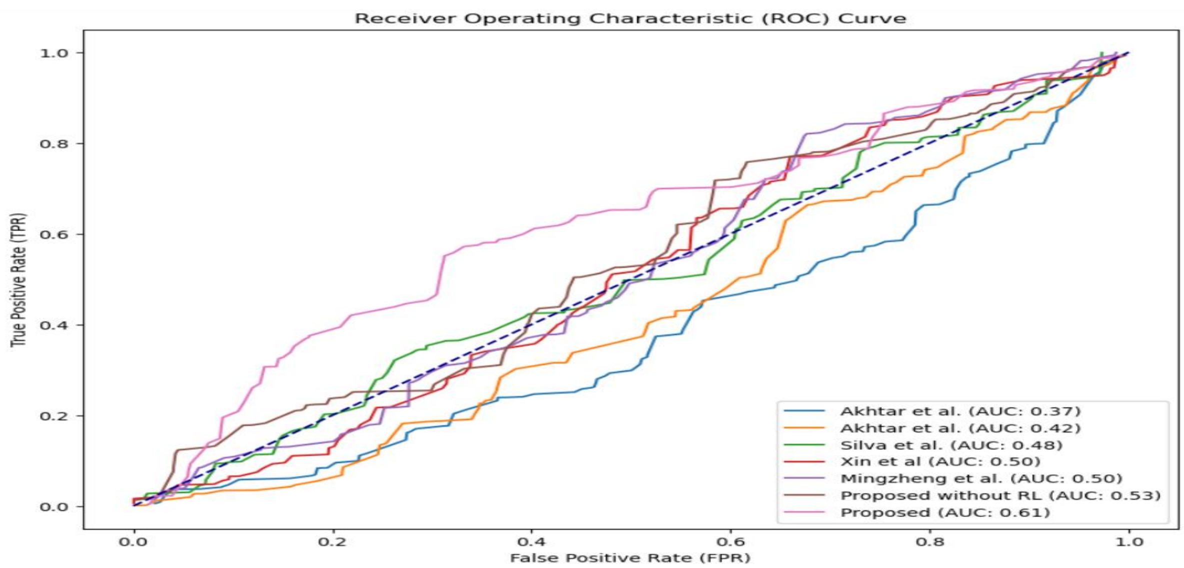


Fig. 5. ROC diagram for the proposed model and other methods. The Blue dashed line represents the ROC curve for a random guess.

2) Impact of the reward function: In SA, the allocation of rewards for both accurate and inaccurate classifications is distributed between majority and minority classes, marked as ±1 and ±λ, correspondingly. The determination of λ is closely associated with the ratio of majority to minority class instances. As this ratio increases, it is anticipated that the ideal λ value will demonstrate a corresponding decrease. To examine the effect of λ on model performance, the model underwent testing with a range of λ values, extending from 0 to 1 in increments of 0.1. During this process, the reward for the majority class remained constant. These experiments and their outcomes are clearly depicted in Fig. 6. When λ is set to 0, the significance afforded to the majority class is effectively negated, while a λ setting of 1 achieves an equilibrium in the treatment of both majority and minority classes. As illustrated in Fig. 6, the model's peak performance is observed at a λ value of 0.7, cutting across all evaluated metrics. This indicates that the most beneficial λ value does not lie at the extremes (0 or 1) but rather at a midpoint. It is crucial to recognize that while it's important to lessen the majority class's dominance by adjusting λ, selecting a value that is too low can adversely affect the model's overall effectiveness. The collected data underlines the significant impact of λ on the performance of the SA model. The ideal λ value is contingent upon the proportion of majority to minority samples, making its careful selection imperative for optimal results. This nuanced approach to λ selection is integral to balancing classification accuracy and ensuring the most advantageous outcomes in SA model applications.

3) Impact of loss function: In addressing the issue of skewed data distribution within SA, the utilization of various conventional methodologies, such as the adaptation of data augmentation techniques and loss functions, is essential. Of these methodologies, the selection of an apt loss function is particularly significant as it can effectively accentuate the importance of the minority class. In this study, a rigorous evaluation of five distinct loss functions is conducted in this study, namely weighted cross-entropy (WCE) [74], balanced cross-entropy (BCE)] [75], dice loss (DL) [76], Tversky loss (TL) [77], and combo Loss (CL) [78]. These were assessed in relation to the proposed model. Notably, the BCE and WCE loss functions ensure unbiased consideration by assigning equivalent weightage to positive and negative samples. In this assortment of functions, the CL function has proven to be an efficient tool for applications wrestling with imbalanced data. This is achieved by its judicious weight distribution; wherein lesser weight is attributed to simpler examples and higher weight to complex ones. This strategic allocation emphasizes the challenging, often underrepresented cases without downplaying the simpler ones. As evidenced in Table III, the CL function, when compared to TL, generates a lower error rate, with the reduction ranging between 15% and 29% across accuracy and F-measure metrics. This remarkable reduction exhibits the CL function's competence in managing imbalanced data more effectively. However, notwithstanding these promising outcomes, it is vital to recognize that the performance of the CL function pales in comparison to RL, trailing behind by a significant 41% margin. This implies that while the CL function can be a feasible option for handling imbalanced data, more sophisticated techniques like RL may offer more precise results.

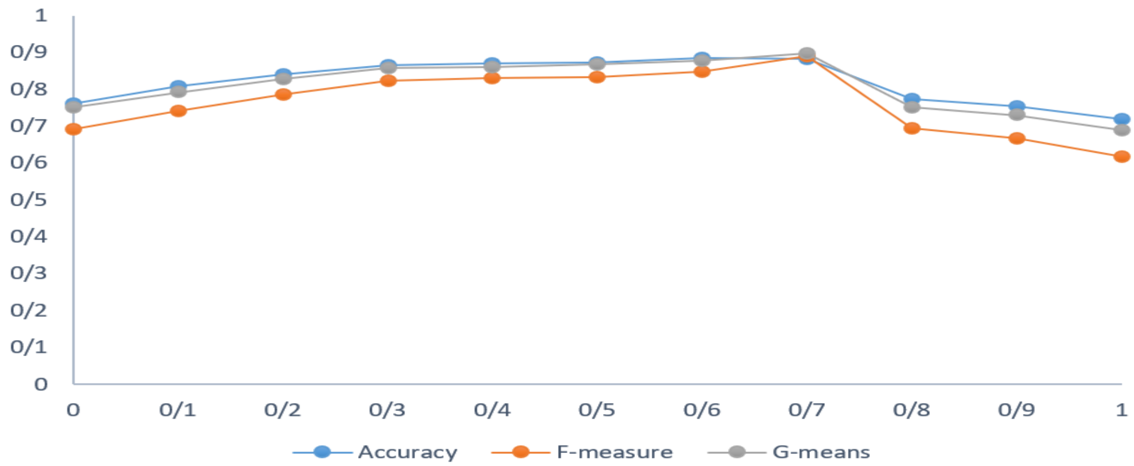


Fig. 6. The performance metrics of the suggested model were graphically represented in relation to the value of λ within the reward function.

TABLE III. RESULTS OF DIFFERENT LOSS FUNCTIONS FOR SA.

Loss function	Accuracy	F-measure	G-means
WCE	0.75±0.03	0.74±0.00	0.76±0.03
BCE	0.80±0.02	0.77±0.01	0.81±0.00
DL	0.81±0.03	0.80±0.01	0.82±0.00
TL	0.83±0.12	0.81±0.04	0.84±0.06
CL	0.86±0.00	0.84±0.04	0.86±0.15

4) Discussion: In the domain of big data and predictive analytics, this groundbreaking study introduces a pioneering approach that combines sentiment analysis from social media with historical stock prices to accurately forecast future stock prices. By harnessing the knowledge exchange platforms on the Internet, the authors offer a fresh perspective on stock market prediction, emphasizing the significant influence of public sentiment on financial market trends. The proposed research presents a novel methodology that analyzes social media by integrating public sentiment, opinions, news, and past stock prices to predict future stock prices. The methodology consists of two main stages. In the initial stage, a SA model is employed with three dilated convolution layers, enabling simultaneous feature extraction and classification. This approach effectively captures the underlying emotions expressed in user-generated content on social media, providing valuable insights into stock market trends. However, SA often faces the challenge of unbalanced classification, which occurs when one class of data is significantly more prevalent than the others. To overcome this challenge, an innovative RL strategy is introduced, treating the task as a sequential decision-making process. The agent receives rewards at each step for accurate classification, with smaller rewards assigned to the majority class compared to the minority class. This approach improves classification accuracy and enables the model to differentiate between different sentiments more effectively. In the second stage, the study incorporates an attention-based LSTM approach, which integrates historical stock prices and the sentiment analysis results obtained in the previous stage to predict future stock prices. The research validates the effectiveness of the proposed model through a series of ablation studies, confirming the positive contributions of the attention-based LSTM and RL components to the overall model performance. The innovative methodology represents a promising advancement in predictive analytics, particularly due to its unique integration of sentiment analysis and historical stock prices. Additionally, the RL strategy for addressing class imbalance adds an intriguing aspect to the field.

Fig. 7 provides a visual representation of the error trajectory observed in the proposed model. It demonstrates the model's progression over 150 training epochs, showcasing a consistent decrease in error. This signifies the continuous improvement of the model's ability to predict stock market trends. The gradual reduction in error rates indicates the convergence of the model towards an optimal solution. Notably, in the final stages of training, the error rates reach an exceptionally low level, such as 0.0000001. This significant decrease underscores the outstanding predictive accuracy achieved by LSTM.

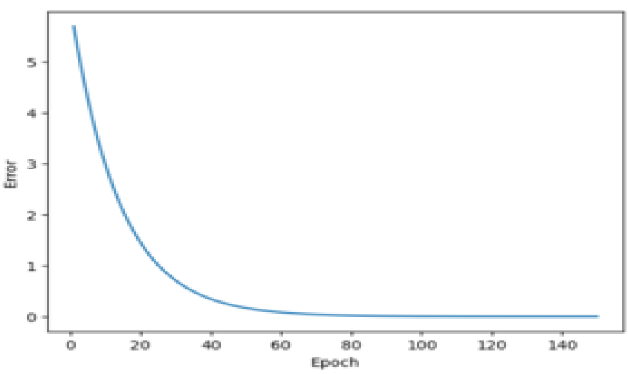


Fig. 7. Diagram of model error during 150 epochs.

Despite the encouraging findings obtained from the research, it is important to acknowledge and address certain limitations that emerged. These limitations not only provide opportunities for further investigation but also indicate areas where improvements can be made in future studies.

- The effectiveness of the proposed methodology relies on the quality and representativeness of the data utilized. However, social media posts and financial news, which serve as the primary data sources, can sometimes be unreliable or inaccurate, posing a risk to the analysis and prediction process [79]. To address this issue, future research can focus on refining data pre-processing techniques to filter out misleading or irrelevant posts, ensuring that sentiment analysis is based on accurate and reflective data [80]. Additionally, there is a need to develop more advanced NLP techniques that can better understand the context and sentiments expressed in social media posts and financial news. Expanding the scope of research, incorporating techniques like sentiment intensity analysis, emotion detection, irony and sarcasm detection, and stance detection can provide a more comprehensive understanding of public sentiment.

- Furthermore, exploring methods to assess the reliability and credibility of data sources, such as establishing a rating system for social media platforms or news outlets, can enhance the accuracy of sentiment analysis [81]. Looking ahead, the potential of machine learning and AI in this field is vast. Future work can explore the integration of other data forms, including audio and video content from different social media platforms, to gain more diverse insights into public sentiment and its impact on stock market trends [82].

- The current methodology employs a fixed reward ratio for the majority and minority classes, which has demonstrated effectiveness in the experiments. However, it may not be optimal for all datasets, particularly those with varying degrees of class imbalance. To address this, future research could explore the dynamic adjustment of the reward function based on the observed class distribution in the data.

- In addition, the proposed approach only considers immediate historical stock prices for prediction. To enhance the predictive capabilities of the model, future studies could investigate the inclusion of longer historical periods or incorporate other relevant factors such as industry trends, economic indicators, or global events. Furthermore, it is worth noting that the proposed method was tested on a limited number of stock markets. To ensure the generalizability of the model, it would be valuable to validate the approach using data from various global stock markets.

- A possible direction for future work could be to incorporate inductive learning elements, allowing the model to generalize from past trends and apply them to future predictions [83]. This extension would enable the model to capture and leverage the underlying patterns and dynamics of the stock market that span across different time periods. By integrating inductive learning, the model can learn from historical data and extract valuable insights that can be applied to forecast future stock prices with greater accuracy. This approach would involve incorporating techniques such as time-series analysis, trend identification, and pattern recognition to identify recurring patterns and trends in the data. By recognizing and understanding these patterns, the model can make more informed predictions about future market behavior [84].

- Additionally, incorporating inductive learning would help the model adapt and adjust its predictions as new data becomes available, ensuring its relevance and effectiveness in dynamic market environments. This approach holds great potential for improving the long-term forecasting capabilities of the model, enabling it to anticipate market trends and make proactive investment decisions. Overall, the incorporation of inductive learning elements represents an exciting avenue for future research, offering the opportunity to enhance the predictive power and robustness of the model in stock market analysis.

- The proposed model primarily focuses on analyzing sentiments from social media platforms. However, future work could look into incorporating various other types of online sources that influence the stock market. Such sources could include financial forums, professional analyst reports, business news websites, and investor behavior analytics from trading platforms [85]. These additional sources of data could provide a more comprehensive and diversified input into the model, leading to potentially higher predictive accuracy. Advanced web scraping techniques and API usage could be employed to harvest data from these varied sources, while machine learning algorithms can be applied to analyze and incorporate this data into the predictive model.

- The current SA model uses a single layer of sentiment classification, which categorizes the sentiment into positive, negative, or neutral. While this works effectively, it may overlook nuanced sentiment gradations that can significantly influence stock prices. Future research could introduce multi-dimensional sentiment analysis that captures not only the polarity of the sentiment but also the intensity and the emotional context. This can involve using techniques like aspect-based sentiment analysis (ABSA) [86], which extracts sentiments related to specific aspects of a topic. For example, a social media post might express positive sentiment about a company's management but a negative sentiment about its new product. Recognizing and capturing these nuanced sentiments can provide a more accurate sentiment representation and lead to improved stock price predictions.

- Considering the potential influence of influential individuals on the stock market, another future research direction could involve the analysis of sentiment expressed by key figures in the business world or financial analysts. Often, the views expressed by these individuals can significantly sway public sentiment and impact the stock market. Using entity recognition techniques, future models can differentiate comments made by these influential figures and assign higher weights to these in the sentiment analysis process.

- While the study provides an innovative approach to predicting stock prices, it is also important to study how the proposed methodology performs under various market conditions. Future work could include stress testing the model under different market scenarios such as bull markets, bear markets, and periods of high volatility. This would provide more insights into the model's robustness and reliability under different circumstances.

- Lastly, the use of deep learning methods in stock price prediction comes with the risk of overfitting, especially when the model becomes too complex [87]. Future research could look into the application of regularization techniques or ensemble methods to mitigate this risk. Additionally, more advanced model interpretability and explainability techniques can be explored to better understand the model's decision-making process and provide more insights to financial analysts and investors. This could further build confidence in the use of AI-based predictive models in financial markets.

V. Conclusion

This research introduces an innovative method for leveraging social media analysis to forecast future trends in stock prices. The proposed methodology encompasses two primary phases. During the initial phase, a SA model was constructed, employing three dilated convolutional layers. These layers were instrumental in simultaneously extracting feature vectors, which were subsequently amalgamated for the purpose of classification. Nonetheless, the SA model encountered a notable hurdle in the form of imbalanced classification. To address this challenge, a RL strategy was devised. This strategy framed the classification task as a series of sequential decision-making events, wherein the agent garnered rewards for each instance of precise classification. To effectively manage the issue of class disparity, the model was designed to allocate reduced rewards for classifications pertaining to the predominant class in contrast to those of the minority class. Additionally, an enhanced DE algorithm was applied for the initialization of the weights in the SA model. In the second phase of the methodology, an attention-based LSTM technique was employed. This phase intricately combined historical stock market data with insights derived from the SA model in the preceding stage. The integration of these two data sources was instrumental in generating more accurate predictions of future stock prices, showcasing the efficacy and potential of this dual-stage approach in the realm of financial forecasting.

The computational challenges inherent in large-scale models such as BERT are significant and widely recognized. While these models boast advanced capabilities, they are also associated with a considerable computational load, largely due to their architecture involving millions of parameters. To effectively address this issue, one promising strategy is the exploration and adoption of more streamlined and efficient variants of the BERT model. In this context, DistilBERT stands out as a leading example, epitomizing this innovative class of models. DistilBERT effectively preserves a majority of the functionality inherent in the original BERT model, while achieving a notable reduction in both the model's size and its computational requirements. This successful blend of high-level performance coupled with increased efficiency highlights the groundbreaking potential of such models. They offer a promising avenue for revolutionizing the computational framework within the realm of large-scale NLP, paving the way for more accessible and sustainable AI solutions in this domain.

References

[1] B. Sun and V. T. Ng, "Analyzing sentimental influence of posts on social networks," in Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2014: IEEE, pp. 546-551.

[2] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," International Journal, vol. 2, no. 6, pp. 282-292, 2012.

[3] F. Z. Xing, E. Cambria, and R. E. Welsch, "Intelligent asset allocation via market sentiment views," ieee ComputatioNal iNtelligeNCe magaziNe, vol. 13, no. 4, pp. 25-34, 2018.

[4] V. Barot, V. Kapadia, and S. Pandya, "QoS enabled IoT based low cost air quality monitoring system with power consumption optimization," Cybernetics and Information Technologies, vol. 20, no. 2, pp. 122-140, 2020.

[5] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, and K. Kotecha, "Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease," PeerJ Computer Science, vol. 7, p. e369, 2021.

[6] M. Soleimani, Z. Forouzanfar, M. Soltani, and M. J. Harandi, "Imbalanced Multiclass Medical Data Classification based on Learning Automata and Neural Network," EAI Endorsed Transactions on AI and Robotics, vol. 2, 2023.

[7] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in International conference on intelligent computing, 2005: Springer, pp. 878-887.

[8] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in Proceedings of workshop on learning from imbalanced datasets, 2003, vol. 126: ICML, pp. 1-7.

[9] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning from imbalanced data sets. Springer, 2018.

[10] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in 2016 international joint conference on neural networks (IJCNN), 2016: IEEE, pp. 4368-4374.

[11] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5375-5384.

[12] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," Applied Intelligence, vol. 50, pp. 2488-2502, 2020.

[13] S. V. Moravvej, M. Joodaki, M. J. M. Kahaki, and M. S. Sartakhti, "A method based on an attention mechanism to measure the similarity of two sentences," in 2021 7th International Conference on Web Research (ICWR), 2021: IEEE, pp. 238-242.

[14] S. V. Moravvej, A. Mirzaei, and M. Safayani, "Biomedical text summarization using conditional generative adversarial network (CGAN)," arXiv preprint arXiv:2110.11870, 2021.

[15] S. Moravvej, M. Maleki Kahaki, M. Salimi Sartakhti, and M. Joodaki, "Efficient GAN-based method for extractive summarization," Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 10, no. 2, pp. 287-298, 2022.

[16] M. Marani, M. Soltani, M. Bahadori, M. Soleimani, and A. Moshayedi, "The Role of Biometric in Banking: A Review," EAI Endorsed Transactions on AI and Robotics, vol. 2, no. 1, 2023.

[17] S. V. Moravvej, S. J. Mousavirad, D. Oliva, and F. Mohammadi, "A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm," arXiv preprint arXiv:2305.02374, 2023.

[18] H. Gharagozlou, J. Mohammadzadeh, A. Bastanfard, and S. S. Ghidary, "RLAS-BIABC: A reinforcement learning-based answer selection using the bert model boosted by an improved ABC algorithm," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[19] T. Eltaeib and A. Mahmood, "Differential evolution: A survey and analysis," Applied Sciences, vol. 8, no. 10, p. 1945, 2018.

[20] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the artificial bee colony (ABC) algorithm in collaboration with the fog nodes in the Internet of Things three-layer architecture," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 509-513.

[21] N. Milosevic, "Equity forecast: Predicting long term stock price movement using machine learning," arXiv preprint arXiv:1603.00751, 2016.

[22] S. Pandya, A. Sur, and K. Kotecha, "Smart epidemic tunnel: IoT-based sensor-fusion assistive technology for COVID-19 disinfection," International Journal of Pervasive Computing and Communications, vol. 18, no. 4, pp. 376-387, 2022.

[23] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in 2013 IEEE 13th International Conference on Data Mining Workshops, 2013: IEEE, pp. 440-444.

[24] M. Athale, S. Nakod, and A. Kumar, "Stock Analysis using Sentiment Analysis and Machine Learning," 2020.

[25] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," International Journal of Scientific and Technology Research, vol. 9, no. 2, pp. 601-609, 2020.

[26] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," Artificial Intelligence Review, vol. 50, no. 1, pp. 49-73, 2018.

[27] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Predictive sentiment analysis of tweets: A stock market application," in Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data: Third International Workshop, HCI-KDD 2013, Held at SouthCHI 2013, Maribor, Slovenia, July 1-3, 2013. Proceedings, 2013: Springer, pp. 77-88.

[28] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, "Stock market predication using a linear regression," in 2017 international conference of electronics, communication and aerospace technology (ICECA), 2017, vol. 2: IEEE, pp. 510-513.

[29] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," Expert systems with applications, vol. 42, no. 1, pp. 259-268, 2015.

[30] C. I. Patel, D. Labana, S. Pandya, K. Modi, H. Ghayvat, and M. Awais, "Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences," Sensors, vol. 20, no. 24, p. 7299, 2020.

[31] A. E. Khedr and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," International Journal of Intelligent Systems and Applications, vol. 9, no. 7, p. 22, 2017.

[32] P. Li and A. Hawbani, "An efficient budget allocation algorithm for multi-channel advertising," in 2018 24th International Conference on Pattern Recognition (ICPR), 2018: IEEE, pp. 886-891.

[33] A. Carosia, G. P. Coelho, and A. Silva, "Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media," Applied Artificial Intelligence, vol. 34, no. 1, pp. 1-19, 2020.

[34] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69-78.

[35] A. Yoshihara, K. Fujikawa, K. Seki, and K. Uehara, "Predicting stock market trends by recurrent deep neural networks," in PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13, 2014: Springer, pp. 759-769.

[36] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," Expert Systems with Applications, vol. 184, p. 115537, 2021.

[37] X. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang, "An innovative neural network approach for stock market prediction," The Journal of Supercomputing, vol. 76, pp. 2098-2118, 2020.

[38] A. Sun, M. Lachanski, and F. J. Fabozzi, "Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction," International Review of Financial Analysis, vol. 48, pp. 272-281, 2016.

[39] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," Journal of Ambient Intelligence and Humanized Computing, pp. 1-24, 2020.

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

[41] M. Bahadori, M. Soltani, M. Soleimani, and M. Bahadori, "Statistical Modeling in Healthcare: Shaping the Future of Medical Research and Healthcare Delivery," in AI and IoT-Based Technologies for Precision Medicine: IGI Global, 2023, pp. 431-446.

[42] K. V. Price, "Differential evolution," Handbook of Optimization: From Classical to Modern Approach, pp. 187-214, 2013.

[43] S. Vakilian, S. V. Moravvej, and A. Fanian, "Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer," in 2021 5th International Conference on Internet of Things and Applications (IoT), 2021: IEEE, pp. 1-5.

[44] M. Arafa, E. A. Sallam, and M. Fahmy, "An enhanced differential evolution optimization algorithm," in 2014 fourth international conference on digital information and communication technology and its applications (DICTAP), 2014: IEEE, pp. 216-225.

[45] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer, and Z. Sobhaninia, "An improved de algorithm to optimise the learning process of a bert-based plagiarism detection model," in 2022 IEEE Congress on Evolutionary Computation (CEC), 2022: IEEE, pp. 1-7.

[46] K. Deb, "A population-based algorithm-generator for real-parameter optimization," Soft Computing, vol. 9, pp. 236-253, 2005.

[47] S. V. Moravvej, M. J. M. Kahaki, M. S. Sartakhti, and A. Mirzaei, "A method based on attention mechanism using bidirectional long-short term memory (BLSTM) for question answering," in 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021: IEEE, pp. 460-464.

[48] M. S. Sartakhti, M. J. M. Kahaki, S. V. Moravvej, M. javadi Joortani, and A. Bagheri, "Persian language model based on BiLSTM model on COVID-19 corpus," in 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021: IEEE, pp. 1-5.

[49] S. V. Moravvej, S. J. Mousavirad, M. H. Moghadam, and M. Saadatmand, "An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes," in Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, 2021: Springer, pp. 690-701.

[50] L. Hong et al., "GAN ‐ LSTM ‐ 3D: An efficient method for lung tumour 3D reconstruction enhanced by attention ‐ based LSTM," CAAI Transactions on Intelligence Technology, 2023.

[51] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," in 2015 IEEE international conference on big data (big data), 2015: IEEE, pp. 2823-2824.

[52] H. Zareiamand, A. Darroudi, I. Mohammadi, S. V. Moravvej, S. Danaei, and R. Alizadehsani, "Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review," Diagnostics, vol. 13, no. 16, p. 2667, 2023.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[54] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[55] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," Journal of machine learning research, vol. 3, no. Aug, pp. 115-143, 2002.

[56] A. Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[57] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," arXiv preprint arXiv:1409.2329, 2014.

[58] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, p. e623, 2021.

[59] S. V. Moravvej et al., "RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights," Contrast Media & Molecular Imaging, vol. 2022, 2022.

[60] K. Zhang, G. Zhong, J. Dong, S. Wang, and Y. Wang, "Stock market prediction based on generative adversarial network," Procedia computer science, vol. 147, pp. 400-406, 2019.

[61] J. Rasheed, A. Jamil, A. A. Hameed, M. Ilyas, A. Özyavaş, and N. Ajlouni, "Improving stock prediction accuracy using cnn and lstm," in 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020: IEEE, pp. 1-5.

[62] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," Neural Computing and Applications, vol. 32, pp. 9713-9729, 2020.

[63] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," Procedia computer science, vol. 167, pp. 599-606, 2020.

[64] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," PeerJ Computer Science, vol. 7, p. e476, 2021.

[65] S. R. Riady, "Stock Price Prediction using Prophet Facebook Algorithm for BBCA and TLKM," International Journal of Advances in Data and Information Systems, vol. 4, no. 1, pp. 1-8, 2023.

[66] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep Transformer model," Expert Systems with Applications, vol. 208, p. 118128, 2022.

[67] S. BL and S. BR, "Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments," Kybernetes, vol. 52, no. 3, pp. 748-773, 2023.

[68] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," Neurocomputing, vol. 398, pp. 247-256, 2020.

[69] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, "Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis," Knowledge-Based Systems, vol. 125, pp. 116-135, 2017.

[70] J. Silva et al., "Algorithm for Detecting Opinion Polarity in Laptop and Restaurant Domains," Procedia Computer Science, vol. 170, pp. 977-982, 2020.

[71] X. Xin, A. Wumaier, Z. Kadeer, and J. He, "SSEMGAT: Syntactic and Semantic Enhanced Multi-Layer Graph Attention Network for Aspect-Level Sentiment Analysis," Applied Sciences, vol. 13, no. 8, p. 5085, 2023.

[72] L. Mingzheng, H. Zelin, L. Jiadong, and L. Wei, "Aspect-level sentiment analysis model fused with GPT and multi-layer attention," in Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), 2023, vol. 12610: SPIE, pp. 279-284.

[73] S. Danaei et al., "Myocarditis Diagnosis: A Method using Mutual Learning-Based ABC and Reinforcement Learning," in 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo), 2022: IEEE, pp. 000265-000270.

[74] Ö. Özdemir and E. B. Sönmez, "Weighted cross-entropy for unbalanced data with application on covid x-ray images," in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020: IEEE, pp. 1-6.

[75] F. Huang, J. Li, and X. Zhu, "Balanced Symmetric Cross Entropy for Large Scale Imbalanced and Noisy Data," arXiv preprint arXiv:2007.01618, 2020.

[76] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," arXiv preprint arXiv:1911.02855, 2019.

[77] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8, 2017: Springer, pp. 379-387.

[78] S. A. Taghanaki et al., "Combo loss: Handling input and output imbalance in multi-organ segmentation," Computerized Medical Imaging and Graphics, vol. 75, pp. 24-33, 2019.

[79] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp. 22-36, 2017.

[80] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," in Proceedings of the Fourth ACM International Conference on AI in Finance, 2023, pp. 349-356.

[81] X. Liu, C. Li, J. L. Nicolau, and M. Han, "The value of rating diversity within multidimensional rating system: Evidence from hotel booking platform," International Journal of Hospitality Management, vol. 110, p. 103434, 2023.

[82] K. Nyakurukwa and Y. Seetharam, "The evolution of studies on social media sentiment in the stock market: Insights from bibliometric analysis," Scientific African, p. e01596, 2023.

[83] A. Modirrousta-Galian, P. A. Higham, and T. Seabrooke, "Effects of inductive learning and gamification on news veracity discernment," Journal of Experimental Psychology: Applied, 2023.

[84] Z. Ding, J. Wu, Z. Li, Y. Ma, and V. Tresp, "Improving Few-Shot Inductive Learning on Temporal Knowledge Graphs using Confidence-Augmented Reinforcement Learning," arXiv preprint arXiv:2304.00613, 2023.

[85] H. Xia, S. Chen, J. Z. Zhang, and Y. Liu, "Superposition effect of online news on fintech platforms," International Journal of Emerging Markets, 2023.

[86] P. Mehra, "Unexpected surprise: Emotion analysis and aspect based sentiment analysis (ABSA) of user generated comments to study behavioral intentions of tourists," Tourism Management Perspectives, vol. 45, p. 101063, 2023.

[87] K. Olorunnimbe and H. Viktor, "Deep learning in the stock market—a systematic survey of practice, backtesting, and applications," Artificial Intelligence Review, vol. 56, no. 3, pp. 2057-2109, 2023.