SentimentLP: unveiling advanced sentiment analysis through Leptotila optimization-based gradient boosting machines

Anitha Merlin Durairaj John Louis[1,2], Vimal Kumar Dhanasekaran[2]
[1]Department of Computer Science, Nehru Arts and Science College, Coimbatore, India
[2]Department of Computer Science, Rathinam College of Arts and Science, Coimbatore, India

Article Info

Article history:

Received Jul 3, 2024
Revised Oct 9, 2024
Accepted Oct 17, 2024

Keywords:

Gradient boosting
Lepotila optimization
Neural network
Online shopping
Sentiment

ABSTRACT

Sentiment analysis is pivotal in extracting insights from textual data, enabling organizations to understand customer opinions, market trends, and brand perception. This study introduces a novel approach, SentimentLP, which integrates Leptotila optimization (LPO) with gradient boosting machines (GBM) for sentiment analysis tasks. The proposed framework leverages LPO's dynamic optimization capabilities to enhance GBM models' performance in sentiment classification. Through iterative refinement and adaptive learning, SentimentLP optimizes feature extraction, model training, and ensemble learning processes, improving sentiment analysis accuracy and efficiency. Results from various evaluation metrics, including precision, recall, classification accuracy, and F-measure, demonstrate the effectiveness of SentimentLP in accurately capturing sentiment expressions in text data. Additionally, the fusion of LPO with GBM ensures scalability, adaptability, and interpretability of sentiment analysis models, making SentimentLP a valuable tool for extracting actionable insights from textual data across diverse domains and applications.

Corresponding Author:

Anitha Merlin Durairaj John Louis
Department of Computer Science, Nehru Arts and Science College
Coimbatore, India
Email: merlin.celestino@gmail.com

1. INTRODUCTION

Online shopping has revolutionized how consumers access goods and services, reshaping traditional retail paradigms worldwide [1]. With the proliferation of e-commerce platforms and digital marketplaces, individuals now enjoy unprecedented convenience and accessibility to various products at their fingertips. The evolution of online shopping over the years has been characterized by a seamless integration of technology, user-centric design, and innovative business models catering to modern consumers' diverse needs and preferences [2]. Online shopping has not only transformed consumer behavior but also propelled the growth of digital economies worldwide, heralding a new era of commerce in the digital age. Online shopping offers supreme convenience, allowing customers to browse and buy products from where they are without traditional store hours or geographical limitations. Moreover, online retailers often provide a more comprehensive selection of products than brick-and-mortar stores, enabling consumers to find what they need quickly.

In natural language processing, sentiment analysis is a pivotal technique, offering insights into the emotional undercurrents embedded within textual data. At its core, sentiment analysis endeavors to decipher the sentiments, opinions, and attitudes expressed within written communication, facilitating a deeper understanding of human sentiment at scale [3]. Stemming from the intersection of linguistics, machine learning, and computational semantics, sentiment analysis has garnered widespread attention across diverse domains, ranging from market research and social media monitoring to customer feedback analysis. By discerning the polarity and intensity of sentiments conveyed in text, sentiment analysis empowers organizations to glean actionable insights, inform strategic decision-making, and tailor communication strategies to resonate with target audiences effectively [4]. The evolution of sentiment analysis techniques, encompassing lexicon-based approaches, machine learning algorithms, and deep learning models, underscores its adaptability and relevance in an ever-changing digital landscape.

Sentiment analysis is pivotal in online shopping, offering valuable insights into consumer perceptions, preferences, and behaviors. Businesses can gauge the overall sentiment towards their products and services by analyzing the sentiments expressed in product reviews, social media discussions, and customer feedback. This enables them to identify areas of strength and areas for improvement, guiding product development, marketing strategies, and customer engagement initiatives [5]. Bio-inspired computing mimics natural processes like evolution, foraging, and swarm intelligence to solve complex problems [6], [7]. In sentiment analysis, it enhances model optimization, feature selection, and classification accuracy by applying algorithms like particle swarm optimization or genetic algorithms. This approach improves handling of nuanced sentiment patterns in large-scale text data [8].

2. LITERATURE REVIEW

Utilizing the sequential structure inherent in financial news data and the memory retention abilities of long short-term memory (LSTM) networks, the financial news LSTM analyst (FNLA) [9] aims to grasp intricate sentiment dynamics. This methodology involves initial preprocessing of financial news articles followed by their input into LSTM layers. These layers are designed to acquire knowledge of temporal dependencies and sentiment trends, allowing FNLA to effectively capture nuanced sentiment fluctuations. “Senti-Prompt Embedding” [10] is an embedding technique for sentiment analysis. It proposes a unique embedding method that integrates sentiment prompts into the embedding model to enhance sentiment classification accuracy.

“Interactive Graph Convolution” [3] is an approach for aspect-based sentiment analysis in Chinese text using an interactive graph convolutional network (GCN) augmented with affective knowledge. It contributes by incorporating affective knowledge into the GCN framework, enabling more effective modeling of sentiment relationships in Chinese text. “Deep-Sentiment with D-RNN” [11] presents a deep learning approach for sentiment analysis using a decision-based recurrent neural network (D-RNN). Processing input sequences using recurrent neural networks (RNNs) and making decisions based on learned representations to predict sentiment labels. “EATN” [12] is a transfer learning framework that adapts to different domains and tasks while minimizing computational costs. The working mechanism involves learning domain-specific representations through adaptive feature extraction and transferring knowledge from a pre-trained model to improve performance on target tasks. “TLSA” [13] is an approach for automated sentiment analysis using a combination of transformer and lexicon-based methods. It involves employing transformer models to capture contextual information from textual data and integrating sentiment lexicons to augment sentiment analysis capabilities. Bio-inspired optimization [14]-[20] lead to better results in major research domains [21].

“Financial News” [22] is an LSTM-based approach for sentiment analysis of financial news articles. It contributes by leveraging LSTM networks to capture temporal dependencies and contextual information in financial text data. The working mechanism involves training LSTM models on historical financial news data to learn patterns and relationships between textual features and sentiment labels. “Visual Sentiment” [23] is a method for enhancing semantic correlations in visual sentiment analysis. It incorporates semantic information into visual features to improve sentiment analysis accuracy. The working mechanism involves extracting visual features from images and leveraging semantic embeddings to enhance the representation of visual content. “ARO-LOGI” [8] operates by first collecting and preprocessing data relevant to sentiment analysis. Subsequently, the Amami rabbit optimization algorithm optimizes the parameters of logistic regression models. This optimization process is guided by the search behavior inspired by the Amami rabbit's survival instincts, ensuring efficient exploration of the solution space. The optimized logistic regression model then analyzes the sentiment of the input data, utilizing the learned patterns to classify sentiment accurately. The optimization mechanism synergizes different operations to yield enhanced performance in all research domains [24]-[26].

2.1. Problem statement and motivation

The problem addressed in sentiment analysis involves accurately classifying textual data into sentiment categories like positive, negative, or neutral. Traditional approaches face difficulties in capturing the nuances of natural language, such as sarcasm, irony, and context dependency, often resulting in inaccuracies. The proposed work focuses on improving sentiment analysis by integrating advanced optimization techniques with machine learning algorithms to enhance accuracy and consistency. This work is motivated by the increasing relevance of sentiment analysis in understanding consumer behavior, particularly in online shopping. By exploring its applications in e-commerce, the article highlights how businesses can use sentiment analysis to refine marketing strategies, improve customer satisfaction, and boost sales. It aims to provide readers with a thorough understanding of the methodologies and algorithms that power sentiment analysis.

2.2. Objectives

The proposed work aims to improve sentiment analysis by developing a robust feature extraction method that captures nuances like sarcasm and context. It seeks to optimize model training for better adaptability and generalization, explore efficient ensemble learning techniques to enhance prediction accuracy, and investigate real-time analysis methods for timely insights. In addition the focus is on ensuring the interpretability and explainability of results, enabling stakeholders to trust and validate the insights for informed decision-making.

3. LEPTOTILA OPTIMIZATION BASED GRADIENT BOOSTING MACHINES

Leptotila optimization (LPO) with gradient boosting machines (GBM) to enhance performance in machine learning tasks. LPO dynamically optimizes GBM parameters, improving feature extraction, model training, and ensemble learning. This synergy boosts scalability, adaptability, and interpretability, making LO-GBM effective for classification, regression, and sentiment analysis tasks.

3.1. Gradient boosting machines

GBM is a machine learning method for sentiment analysis uses decision trees to enhance classification by eradicating previous mistakes. In order to improve model performance, it concentrates on cases of misclassification and applies gradient-based optimization. GBM addresses missing values, processes both category and numerical data, and enables hyperparameter adjustment for increased accuracy. Because of its boosting strategy, which enhances weak areas, sentiment analysis in applications such as social media and product evaluations can be done with great effectiveness.

3.1.1. Data collection

This pivotal phase lays the groundwork for subsequent stages by procuring a comprehensive corpus of textual data. The collected dataset serves as the bedrock upon which the GBM model will be trained to discern sentiment expressions. The essence of data collection can be encapsulated in (1):

$$D = \{d_1, d_2, \dots, d_N\} \tag{1}$$

where D represents the dataset comprising N textual documents d_i. Various sources, including social media, product evaluations, and consumer feedback, contribute to each document's unique expressions of feeling.

Moving forward, it's imperative to preprocess the collected data to facilitate the subsequent stages of sentiment analysis. This involves a series of operations to clean and standardize the textual content. The preprocessing stage can be represented mathematically in (2):

$$Preprocessed_i = Preprocess(d_i) \tag{2}$$

where $Preprocess(d_i)$ denotes the preprocessing function applied to each document d_i, resulting in a preprocessed version denoted by $Preprocessed_i$.

3.1.2. Feature extraction

A prevalent technique for feature extraction in sentiment analysis is the bag-of-words model, which represents each document as a vector and indicates the frequency of occurrence of each word in a predefined vocabulary. Mathematically, the bag-of-words representation can be expressed in (3):

$$X_{ij} = Count\left(W_j, D_i\right) \tag{3}$$

where, X_{ij} denotes the count of word W_j in document D_i, and $Count\left(W_j, D_i\right)$ computes the frequency of occurrence of the word W_j in document D_i.

Another prevalent approach for feature extraction is term frequency-inverse document frequency (TF-IDF), which considers the frequency of a word in a document and weighs it based on its rarity across the entire corpus. The TF-IDF representation can be expressed using (4):

$$TFIDF_{ij} = TF\left(W_j, D_i\right) \times IDF\left(W_j\right) \tag{4}$$

where $TF\left(W_j, D_i\right)$ denotes the term frequency of word W_j in document D_i, and $IDF\left(W_j\right)$ represents the inverse document frequency of word W_j across the entire dataset.

3.1.3. Model training

Building a network of decision trees, with each tree taught to fix the mistakes of the ones before it, is the backbone of model training. The iterative process of building decision trees in GBM can be represented in (5):

$$F_m(x) = F_{m-1}(x) + \lambda . h_m(x) \tag{5}$$

where $F_m(x)$ denotes the prediction made by the ensemble of m decision trees for input x, $F_{m-1}(x)$ represents the prediction made by the ensemble of $(m-1)$ trees, λ denotes the learning rate controlling the contribution of each tree, and $h_m(x)$ represents the prediction made by the m^{th}decision tree.

While training the model, it will compute the gradient of a loss function related to the predictions the current ensemble has made. This gradient guide the subsequent decision tree to focus on areas where the model performs poorly, thereby improving the predictive capability of the ensemble. The gradient calculation process can be mathematically expressed with (6):

$$\nabla L(F_{m-1}(x), y) = -[y - F_{m-1}(x)] \tag{6}$$

where, $\nabla L(F_{m-1}(x), y)$ denotes the gradient of the loss function concerning the predictions made by the ensemble $F_{m-1}(x)$ for input x and true label y. The loss function calculates the degree to which actual labels deviate from predictions.

To minimize the prediction errors and enhance model performance, each subsequent decision tree is constructed to complement the strengths and weaknesses of the existing ensemble. The construction of a new decision tree in GBM can be expressed in (7):

$$h_m(x) = argmin_h \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + h(x_i)\right) \tag{7}$$

where, $h_m(x)$ represents the prediction made by the m^{th} decision tree for input x, $argmin_h$ denotes the function that finds the decision tree h minimizing the sum of the loss function L for all training instances (x_i, y_i), y_i represents the true label for training instance x_i, and $F_{m-1}(x_i)$ represents the prediction made by the ensemble of $(m-1)$ trees for training instance x_i.

3.1.4. Tree construction

Tree construction in GBM is guided by the gradients computed during the gradient calculation step. These gradients provide directional signals, enabling the algorithm to focus on areas where the model exhibits significant prediction errors. The construction of a new decision tree process can be represented using (8):

$$h_m(x) = argmin_h \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + h(x_i)\right) \tag{8}$$

where $h_m(x)$ represents the prediction made by the m^{th} decision tree for input x, $argmin_h$ denotes the function that finds the decision tree h minimizing the sum of the loss function L for all training instances (x_i, y_i), y_i represents the true label for training instance x_i, and $F_{m-1}(x_i)$ represents the prediction made by the ensemble of $(m-1)$ trees for training instance x_i.

3.1.5. Ensemble update

The ensemble update process in GBM involves adding the predictions made by the newly constructed decision tree to the projections made by the existing ensemble. This can be represented using (5). This iterative refinement process gradually improves the model's predictive accuracy and enhances its ability to discern sentiment expressions effectively. The ensemble update process in GBM involves adjusting the learning rate parameter to control the contribution of each newly added decision tree to the ensemble. The learning rate adjustment can be mathematically represented in (9):

$$\lambda = \frac{learning\_rate}{1+tree\_index} \tag{9}$$

where λ denotes the learning rate, $learning\_rate$ represents the initial learning rate parameter and $tree\_index$ denotes the index of the newly added decision tree.

3.1.6. Model evaluation

Comparing the trained GBM model's predictions with the dataset's actual labels is standard practice when evaluating a model. Metrics for evaluation like recall, accuracy, precision, and F-score can be mathematically depicted in (10)-(13):

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{10}$$

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \tag{11}$$

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \tag{12}$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{13}$$

where accuracy refers to the model's correct prediction percentage, precision to the proportion of accurately predicted positive instances relative to the total number of positive cases predicted, recall to the proportion of accurately predicted positive instances relative to the total number of actual positive instances, and F1-score, a balanced measure of model performance, is obtained by harmonically averaging recall and precision.

3.2. Leptotila optimization

LPO is inspired by the characteristics of the Leptotila, a genus of birds known for their efficient foraging behaviour and ability to adapt to varying environmental conditions. The optimization algorithm mimics the following characteristics to search for optimal solutions to optimization problems efficiently.

3.2.1. Initialization

The initialization phase sets the foundation by creating an initial set of candidate solutions within the search space. Random initialization uses the search space to create candidate solutions at random. From a mathematical perspective, it is represented using (14):

$$X^0 = [x_1^0, x_2^0, \dots, x_N^0] \tag{14}$$

where X^0 denotes the initial population of size N, and x_i^0 represents the i^{th} randomly generated candidate solution. Latin hypercube sampling (LHS) is a more sophisticated initialization technique that ensures a more uniform search space coverage. It divides each dimension of the search space into N equal intervals and randomly samples one point from each interval for each dimension. It can be mathematically represented in (19). where X^0 represents the initial population and x_i^0denotes the i^{th}candidate solution obtained through LHS.

3.2.2. Evaluation

The fitness function quantifies the quality of each candidate solution by evaluating its performance in solving the optimization problem. The fitness function $f(X^t)$ can be represented mathematically in (15):

$$f(X^t) = [f(x_1^t), f(x_2^t), \dots, f(x_N^t)] \tag{15}$$

where X^t represents the population at iteration t, and $f(x_I^t)$ denotes the fitness value of the i^{th} candidate solution.

The objective function defines the optimization problem and quantitatively measures how well a candidate solution satisfies the problem's objectives. This objective function $F(x)$ can be expressed mathematically in (16):

$$F(x) = F(x_1, x_2, \dots, x_d) \tag{16}$$

where x represents a candidate solution and $F(x)$ quantifies the objective value associated with solution x.

Pareto dominance is applied to evaluate and prioritize potential solutions in multi-objective optimization issues. The Pareto dominance can be represented using (17):

$$x \prec y \tag{17}$$

where x Pareto dominates y, through the evaluation step, LPO assigns a fitness value to each candidate solution, allowing the algorithm to distinguish between promising and unpromising solutions.

3.2.3. Exploration

Exploration is crucial for navigating the search space and discovering new candidate solutions. Mutation is a common exploration technique that introduces random changes to selected candidate solutions, thereby generating new solutions with potentially different characteristics; the mutation operation can be represented using (18):

$$x_i^{t+1} = x_i^t + \Delta x_i \tag{18}$$

where x_i^{t+1} represents the mutated solution, x_i^t denotes the selected candidate solution at iteration t, and Δx_i represents the mutation vector.

Crossover is another exploration technique commonly used in evolutionary algorithms, where two or more selected candidate solutions exchange information to generate new solutions. This crossover operation can be expressed in (19):

$$x_i^{t+1} = \alpha . x_i^t + (1 - \alpha) . x_j^t \tag{19}$$

where x_i^{t+1} represents the offspring solution, x_i^t and x_j^t denote the selected parent solutions, and α represents the crossover coefficient.

3.2.4. Exploitation

Local search is a common exploitation technique that focuses on refining promising candidate solutions by iteratively exploring their neighbourhoods in the solution space. Local search can be represented in (20):

$$x_i^{t+1} = x_i^t + \Delta x_i \tag{20}$$

where x_i^{t+1} represents the updated solution, x_i^t denotes the selected candidate solution at iteration t, and Δx_i represents the change applied to the solution based on local search. One popular optimization method is gradient descent, which uses the objective function's gradient information to improve candidate solutions until they reach the steepest descent repeatedly; this process is represented in mathematical form in (21):

$$x_i^{t+1} = x_i^t + \eta \nabla F(x_i^t) \tag{21}$$

where x_i^{t+1} represents the updated solution, x_i^t denotes the selected candidate solution at iteration t, η represents the step size (learning rate), and $\nabla F(x_i^t)$ represents the gradient of the objective function concerning x_i^t.

Simulated annealing is a probabilistic optimization algorithm that balances exploration and exploitation by allowing occasional uphill moves to escape local optima. In (22) depicts the probability of accepting a worse solution.

$$P(\Delta E, T) = e^{-\frac{\Delta E}{T}} \tag{22}$$

where ΔE represents the change in objective function value and T represents the temperature parameter controlling the likelihood of accepting worse solutions.

3.2.5. Adaptive adjustment

Gradient descent and other optimization methods use the learning rate to regulate the increment of parameters. Adapting the learning rate based on the optimization progress can improve convergence and stability. The adaptive learning rate can be represented mathematically in (23):

$$\eta_{t+1} = \eta_t . \frac{Recent\ Improvement}{Previous\ Improvement} \quad (23)$$

where η_{t+1} represents the updated learning rate at iteration $t+1$, η_t denotes the current learning rate at iteration t, and $Recent\ Improvement$ and $Previous\ Improvement$ represent the recent and previous improvements in the objective function value, respectively.

Balancing exploration and exploitation is key in optimization. Adaptive methods adjust focus between exploring new solutions and refining existing ones. The exploration factor manages this balance, while population size affects convergence speed and diversity. Adapting the exploration rate based on the optimization progress can improve exploration efficiency in algorithms that employ exploration strategies such as mutation or crossover. The exploration rate adaptation can be represented in (24):

$$\epsilon_{t+1} = \epsilon_t . \frac{Recent\ Improvement}{Previous\ Improvement} \quad (24)$$

where ϵ_{t+1} represents the updated exploration rate at iteration $t+1$, ϵ_t denotes the current exploration rate at iteration t, and $Recent\ Improvement$ and $Previous\ Improvement$ represent the recent and previous improvements in the objective function value, respectively.

3.2.6. Communication

In LPO, promising candidate solutions with high fitness values can be shared among the population to guide other solutions towards regions of the solution space with high potential for improvement.

$$X^{t+1} = X^t + \Delta X \quad (25)$$

In (25), where X^{t+1} represents the updated population at iteration $t+1$, X^t denotes the current population at iteration t, and ΔX represents the information shared among promising solutions. Maintaining diversity within the population is essential for preventing premature convergence and promoting the exploration. Communication mechanisms can encourage diversity by exchanging solutions representing diverse solutions or searching directions.

$$Distance(x_i, x_j) \leq Threshold \quad (26)$$

In (26) where, distance (x_i, x_j) represents the distance between candidate solutions x_i and x_j and $Threshold$ represents the maximum allowable distance. The best-performing solutions within the population can serve as sources of valuable information for guiding other solutions toward promising regions of the solution space. This information sharing via best solutions can be represented mathematically in (27):

$$IX^{t+1} = X^t + \Delta X \quad (27)$$

where IX^{t+1} represents the updated information of the population at iteration ready to share $t+1$, X^t denotes the current population at iteration t, and ΔX represents the information shared among the best solutions.

3.2.7. Iterative improvement

Candidate solutions are iteratively updated based on various optimization techniques such as local search, gradient descent, or evolutionary operators. The iterative solution update is shown mathematically in (28):

$$X^{t+1} = Update(x_i^t) \quad (28)$$

where x_i^{t+1} represents the updated solution at iteration $t+1$ and x_i^t denotes the current solution at iteration t.

The objective function value serves as a metric to assess the quality of candidate solutions. Through iterative improvement, the objective function value is expected to decrease over successive iterations as the optimization process converges towards optimal solutions. The objective function improvement can be mathematically represented in (29):

$$F(x^{t+1}) \leq F(x') \quad (29)$$

where $F(x^{t+1})$ represents the objective function value of the updated solution at iteration $t+1$ and $F(x')$ denotes the objective function value of the current solution at iteration t.

Convergence criteria determine when the optimization process converges sufficiently to a satisfactory solution. A common criterion for convergence is when the optimization process reaches a plateau when the objective function value improves by a certain threshold or when the maximum number of iterations is reached. Convergence criteria can be represented in (30):

$$Convergence = \begin{cases} True, if\ convergence\ criteria\ are\ met \\ False, otherwise \end{cases} \tag{30}$$

Iterative improvement strategies encompass a variety of optimization techniques aimed at enhancing candidate solutions over successive iterations. Finally, the termination phase determines when the optimization process should stop based on predefined criteria. The algorithm terminates once this limit is touched, regardless of whether the optimization has converged to an optimal solution.

3.3. SentimentLP-the fusion of Leptotila optimization with gradient boosting machines for sentiment analysis

Sentiment analysis is used in various fields like marketing and social media, involves identifying opinions or sentiments from text data. SentimentLP combines the strengths of the LPO algorithm with GBM to enhance this process. LPO, inspired by the foraging behavior of Leptotila birds, refines solutions by exploring and optimizing them efficiently. GBM, an ensemble learning method, builds strong predictive models by combining several weaker models and using gradient descent to progressively correct mistakes, improving accuracy in sentiment analysis.

3.3.1. Integration of Leptotila optimization with gradient boosting machine for sentiment analysis

The integration of LPO with GBM for sentiment analysis combines LPO's dynamic optimization with GBM's predictive strength. This fusion named SentimentLP, enhances the accuracy and efficiency of sentiment classification by optimizing model parameters and improving prediction reliability.

a. Data collection and preprocessing: SentimentLP begins with collecting and preprocessing the text data, removing noise and irrelevant information to ensure high-quality input for the subsequent steps.
b. Feature extraction: SentimentLP extracts relevant features from the preprocessed text data. This step involves techniques such as bag-of-words, TF-IDF, or word embeddings to represent the textual information in a format suitable for machine learning algorithms.
c. Model training with GBM: LPO dynamically adjusts the hyperparameters of the GBM model, such as learning rate, tree depth, and number of estimators, to optimize its performance. Through iterative improvement, LPO guides the training process of GBM, enhancing its ability to capture complex patterns and relationships in sentiment data.
d. Gradient calculation: during model training, SentimentLP computes gradients to update the parameters of the GBM model. To improve the model’s prediction performance and minimize the loss function, these gradients show the direction and size of the needed modifications.
e. Tree construction: GBM builds decision trees intending to improve upon earlier mistakes. SentimentLP guides the construction of these trees, ensuring that they capture critical features and patterns relevant to sentiment analysis.
f. Ensemble update: SentimentLP updates the ensemble of trees, combining their predictions to make more accurate sentiment predictions. This ensemble update process involves adjusting the weights of individual trees based on their performance and contribution to the overall model.
g. Model evaluation: SentimentLP checks its efficacy with suitable measures, including F1-score, recall, accuracy, and precision. This test verifies that the model successfully interprets the meaning of the text data.
h. Model deployment: SentimentLP deploys the trained GBM model for sentiment analysis tasks in real-world applications. This deployment phase integrates the model into existing systems or platforms, allowing users to efficiently analyze sentiment in text data.

The fusion of LPO with GBM in SentimentLP offers a robust and efficient approach to sentiment analysis tasks.

4. ABOUT DATASET

The Amazon Review Data (2018) of the home and kitchen category is a rich and expansive dataset, comprising 6,898,955 reviews. This dataset is valuable for researchers, analysts, and businesses seeking insights into consumer preferences, sentiments, and trends within the home and kitchen products market. With its vast scale and granularity, this dataset enables comprehensive analysis of various aspects, including product ratings, review sentiments, and customer feedback. By delving into this dataset, analysts can uncover patterns, correlations, and outliers that may inform strategic decision-making, product development, and marketing initiatives within the home and kitchen industry. The availability of a smaller per-category dataset facilitates focused analysis and experimentation, allowing researchers to explore specific subcategories or product niches in greater detail. Overall, the Amazon Review Data (2018) for the home and kitchen category is a valuable asset for advancing research and innovation in e-commerce, consumer behavior, and sentiment analysis.

5. RESULTS AND DISCUSSION

5.1. Precision and recall analysis

Precision represents the proportion of true positive predictions among all positive predictions made by the model, indicating the accuracy of positive predictions. Recall, also known as sensitivity, measures the proportion of true positives correctly identified by the model among all actual positive instances in the dataset, reflecting the model's ability to capture all positive instances. Figure 1 provides the precision and recall analysis. Analyzing the precision and recall metrics for the FNLA, ARO-LOGI, and LO-GBM models reveals insights into their classification performance.

The FNLA model demonstrates a precision of 69.9093% and a recall of 71.163%, indicating that around 69.91% of the positive predictions made by the model are accurate, and it correctly identifies approximately 71.16% of all actual positive instances. Meanwhile, the ARO-LOGI model exhibits higher precision and recall values, with precision at 81.7006% and recall at 78.877%. This suggests that the ARO-LOGI model achieves better accuracy in its positive predictions and captures a significant portion of the actual positive instances. Similarly, the LO-GBM model showcases superior precision and recall metrics, with precision recorded at 89.5753% and recall at 90.691%. These results indicate that the LO-GBM model outperforms the FNLA and ARO-LOGI models in accuracy and sensitivity, demonstrating its effectiveness in correctly identifying positive instances while minimizing false positives. The outcome of precision and recall is illustrated in Figure 1. The LO-GBM model outperforms FNLA and ARO-LOGI models regarding precision and recall, showcasing its superior performance in sentiment analysis tasks.

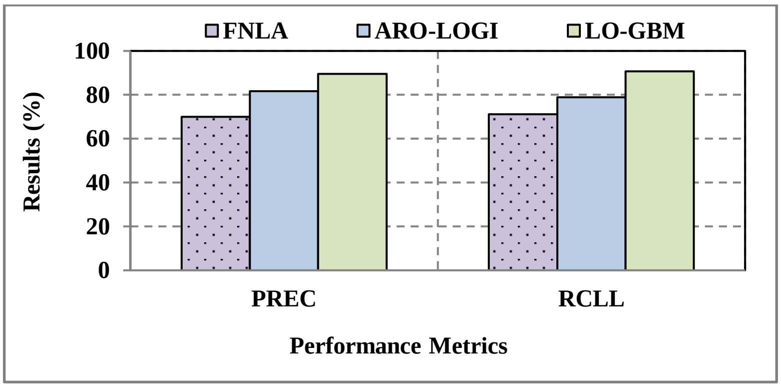

Figure 1. Precision and recall

5.2. Classification accuracy and F-measure analysis

Classification accuracy measures the proportion of correctly classified instances in the dataset. At the same time, F-measure combines precision and recall to provide a metric that balances false positives and negatives. Figure 2 provides the classification accuracy and F-measure analysis. The classification accuracy values for FNLA, ARO-LOGI, and LO-GBM are 69.3970%, 80.2680%, and 89.8280% respectively. This indicates that the LO-GBM model achieves the highest classification accuracy among the three models, with an accuracy of 89.8280%. A higher classification accuracy suggests that the model effectively predicts sentiment labels, resulting in more correct classifications than the total number of instances.

The F-measure values for FNLA, ARO-LOGI, and LO-GBM are 70.5307%, 80.2641%, and 90.1298%, respectively. Similar to classification accuracy, the LO-GBM model also outperforms the other models in terms of F-measure. With an F-measure of 90.1298%, the LO-GBM model achieves the highest balance between precision and recall, indicating a robust performance in sentiment analysis tasks. F-measure provides a more comprehensive evaluation of the model's performance by considering both false positives and negatives.

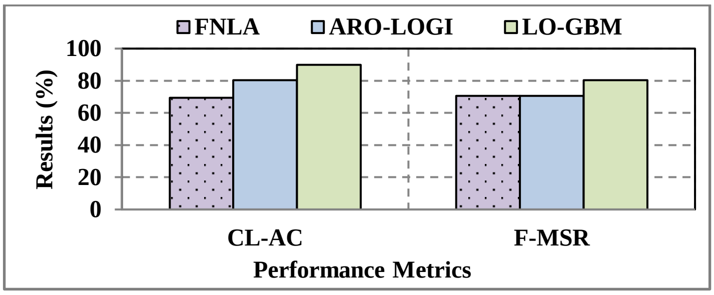

Figure 2. Classification accuracy and F-measure

6. CONCLUSION

The LO-GBM classification algorithm within SentimentLP achieved a remarkable classification accuracy of 89.828% and an F-measure of 80.264%, demonstrating its efficacy in accurately discerning sentiment expressions from textual data. This improvement can be attributed to the synergistic combination of LPO's dynamic optimization capabilities and GBM's ensemble learning approach. The integration of LPO with GBM enhances model performance and ensures scalability, adaptability, and interpretability across diverse domains. SentimentLP's iterative refinement and adaptive learning processes, driven by LPO, optimize feature extraction, model training, and ensemble learning, contributing to its robust performance. These findings underscore SentimentLP's potential as a valuable tool for organizations seeking to extract actionable insights from textual data. By providing a deeper understanding of customer opinions, market trends, and brand perceptions, SentimentLP empowers decision-makers to make informed choices in various domains. Future research can further explore its real-world applications and address remaining challenges, ensuring continued advancements in sentiment analysis. SentimentLP represents a significant leap forward in sentiment analysis, offering a comprehensive solution for extracting valuable insights from textual data across diverse applications.

REFERENCES

[1] J. Ye *et al.*, “Sentiment-aware multimodal pre-training for multimodal sentiment analysis,” *Knowledge-Based Systems*, vol. 258, p. 110021, 2022, doi: 10.1016/j.knosys.2022.110021.

[2] G. Chao, J. Liu, M. Wang, and D. Chu, “Data augmentation for sentiment classification with semantic preservation and diversity,” *Knowledge-Based Systems*, vol. 280, p. 111038, 2023, doi: 10.1016/j.knosys.2023.111038.

[3] Q. Yang, Z. Kadeer, W. Gu, W. Sun, and A. Wumaier, “Affective Knowledge Augmented Interactive Graph Convolutional Network for Chinese-Oriented Aspect-Based Sentiment Analysis,” *IEEE Access*, vol. 10, pp. 130686–130698, 2022, doi: 10.1109/ACCESS.2022.3228299.

[4] H. Wang, C. Ren, and Z. Yu, “Multimodal sentiment analysis based on multiple attention,” *Engineering Applications of Artificial Intelligence*, vol. 140, p. 109731, 2025, doi: 10.1016/j.engappai.2024.109731

[5] P. Menakadevi and J. Ramkumar, “Robust Optimization Based Extreme Learning Machine for Sentiment Analysis in Big Data,” *2022 International Conference on Advanced Computing Technologies and Applications*, *ICACTA 2022*, 2022, pp. 1–5, doi: 10.1109/ICACTA54488.2022.9753203.

[6] M. P. Swapna and J. Ramkumar, “Multiple Memory Image Instances Stratagem to Detect Fileless Malware,” *International Conference on Advancements in Smart Computing and Information Security*, Cham: Springer Nature Switzerland, 2024, pp. 131–140, doi: 10.1007/978-3-031-59100-6_11.

[7] R. Karthikeyan and R. Vadivel, “Proficient Dazzling Crow Optimization Routing Protocol (PDCORP) for Effective Energy Administration in Wireless Sensor Networks,” in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 2023, pp. 1–6, doi: 10.1109/ELEXCOM58812.2023.10370559.

[8] D. J. A. Merlin and D. V. Kumar, “Exploring Aro-Logisent: Delving Into Data Collection For Advanced Sentiment Analysis With Amami Rabbit Optimization-Based Logistic Regression,” *Journal of Theoretical and Applied Information Technology*, vol. 15, no. 9, pp. 3730–3755, 2024.

[9] A. Sharaff, T. R. Chowdhury, and S. Bhandarkar, “LSTM based Sentiment Analysis of Financial News,” *SN Computer Science*, vol. 4, no. 5, p. 584, 2023, doi: 10.1007/s42979-023-02018-2.

[10] J. Kim and Y. Ko, “SPACE: Senti-Prompt As Classifying Embedding for sentiment analysis,” *Pattern Recognition Letters*, vol. 180, pp. 62–67, 2024, doi: 10.1016/j.patrec.2024.02.022.

[11] P. Durga and D. Godavarthi, “Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN),” *IEEE Access*, vol. 11, pp. 108433–108447, 2023, doi: 10.1109/ACCESS.2023.3320738.

[12] K. Zhang *et al.*, “EATN: An Efficient Adaptive Transfer Network for Aspect-Level Sentiment Analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 377–389, 2023, doi: 10.1109/TKDE.2021.3075238.

[13] X. Zhao and C.-W. Wong, “Automated measures of sentiment via transformer- and lexicon-based sentiment analysis (TLSA),” *Journal of Computational Social Science*, vol. 7, no. 1, pp. 145–170, 2024, doi: 10.1007/s42001-023-00233-8.

[14] J. Ramkumar and R. Vadivel, “Improved Wolf prey inspired protocol for routing in cognitive radio Ad Hoc networks,” *International Journal of Computer Networks and Applications*, vol. 7, no. 5, pp. 126–136, 2020, doi: 10.22247/ijcna/2020/202977.

[15] J. Ramkumar and R. Vadivel, “Whale optimization routing protocol for minimizing energy consumption in cognitive radio wireless sensor network,” *International Journal of Computer Networks and Applications*, vol. 8, no. 4, pp. 455–464, 2021, doi: 10.22247/ijcna/2021/209711.

[16] R. Karthikeyan and R. Vadivel, “Boosted Mutated Corona Virus Optimization Routing Protocol (BMCVORP) for Reliable Data Transmission with Efficient Energy Utilization,” *Wireless Pers. Commun.*, vol. 135, no. 4, pp. 2281–2301, 2024.

[17] R. Jaganathan, S. Mehta, and R. Krishan, *Intelligent Decision Making Through Bio-Inspired Optimization*, IGI Global, 2024, doi: 10.4018/979-8-3693-2073-0.

[18] R. Jaganathan, S. Mehta, and R. Krishan, *Bio-Inspired intelligence for smart decision-making*, IGI Global, 2024, doi: 10.4018/9798369352762.

[19] M. P. Swapna, J. Ramkumar, and R. Karthikeyan, “Energy-Aware Reliable Routing with Blockchain Security for Heterogeneous Wireless Sensor Networks,” in *International Conference on Advances in Information Communication Technology & Computing*, Shymkent, 2024, vol. 1075, pp. 713-723, doi: 10.1007/978-981-97-6106-7_43.

[20] N. K. Ojha, A. Pandita, and J. Ramkumar, “Cyber security challenges and dark side of AI: Review and current status,” in *Demystifying the Dark Side of AI in Business*, 2024, pp. 117–137, doi: 10.4018/979-8-3693-0724-3.ch007.

[21] R. Jaganathan and V. Ramasamy, “Performance modeling of bio-inspired routing protocols in Cognitive Radio Ad Hoc Network to reduce end-to-end delay,” *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 221–231, 2019, doi: 10.22266/IJIES2019.0228.22.

[22] J. Maqbool, P. Aggarwal, R. Kaur, A. Mittal, and I. A. Ganaie, “Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach,” *Procedia Computer Science*, vol. 218, pp. 1067–1078, 2023, doi: 10.1016/j.procs.2023.01.086.

[23] H. Zhang, Y. Liu, Z. Xiong, Z. Wu, and D. Xu, “Visual sentiment analysis with semantic correlation enhancement,” *Complex & Intelligent Systems*, vol. 10, pp. 2869–2881, 2023, doi: 10.1007/s40747-023-01296-w.

[24] S. P. Geetha, N. M. S. Sundari, J. Ramkumar, and R. Karthikeyan, “Energy Efficient Routing in Quantum Flying Ad Hoc Network (Q-FANET) Using Mamdani Fuzzy Inference Enhanced Dijkstra’s Algorithm (MFI-EDA),” *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 9, pp. 3708–3724, 2024.

[25] G. M. Balaji and K. Vadivazhagan, “Empowering Online Shopping Sentiment Analysis Using Tenacious Artificial Bee Colony Inspired Taylor Series-Based Gaussian Mixture Model (TABC-TSGMM),” *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 1, pp. 236–257, 2024.

[26] J. Ramkumar, R. Karthikeyan, and M. Lingaraj, “Optimizing IoT-Based Quantum Wireless Sensor Networks Using NM-TEEN Fusion of Energy Efficiency and Systematic Governance,” *International Conference on Power Engineering and Intelligent Systems (PEIS)*. Singapore: Springer, vol. 1246, pp. 141–153, 2025, doi: 10.1007/978-981-97-6710-6_12.

BIOGRAPHIES OF AUTHORS

Anitha Merlin Durairaj John Louis completed his Master of Computer Applications at VLB Janakiammal College of Arts and Science, affiliated with Bharathiar University, India, in 2002. She then obtained his M.Phil. in Computer Science from VLB Janakiammal College of Arts and Science, affiliated with Bharathiar University, in 2007. Subsequently, she is currently pursuving Ph.D. in Nehru Arts and Science College, Coimbatote, India, amassing 11 years of teaching experience. Currently serving as an Assistant Professor (SG) in Department of Computer Science. Nehru Arts and Science College, Coimbatore, she has Published 2 Journals in Scopus and Published 5 Books. She holds one has published five Indian patents. She can be contacted at email: merlin.celestino@gmail.com.

Dr. Vimal Kumar Dhanasekaran completed his Master of Computer Applications at K.S. Rangasamy College of Technology, affiliated with Periyar University, India, in 2002. He then obtained his M.Phil. in Computer Science from Kongu Arts and Science College, affiliated with Bharathiar University, in 2007. Subsequently, he was awarded a Ph.D. from Anna University in 2014, amassing 19 years of teaching experience. Currently serving as an Associate Professor and Head of the Department of Computer Science at Rathinam College of Arts and Science, Coimbatore, Tamil Nadu, India, he is recognized as an approved supervisor at Bharathiar University. He is actively mentoring Three Ph.D. scholars, of whom one has submitted the synopsis. Notably, under his guidance, five Ph.D. scholars and four M.Phil. scholars have been awarded their degrees. With 61 articles published in international journals, including 20 indexed in Scopus and four in Web of Science, he has made significant research contributions. He holds one Australian patent and has published five Indian patents, showcasing his expertise in data mining, networking, IoT, software engineering, mobile computing, and image processing. Recognized for his exceptional work, he has received various awards, including the Best Faculty Award, Best Scientist Award, Best Lecturer Award, Best Social Worker Award, and Star of Hope Award. He has served as Chief Guest and Resource Person in numerous programs organized by educational institutions, delivering lectures at national and international conferences. Actively engaged in professional circles, he is a member of prestigious professional bodies and serves as a journal reviewer. He can be contacted at email: drvimalcs@gmail.com.