The Scientific Temper (2024) Vol. 15 (3): 2623-2633 E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.30 https://scientifictemper.com/

RESEARCH ARTICLE

MOHCOA: Multi-objective hermit crab optimization algorithm for feature selection in sentiment analysis of Covid-19 Twitter datasets

A. Sathya[*], **M. S. Mythili**

Abstract

The COVID-19 pandemic has led to a flood of data on Twitter, making it crucial to analyze public opinion. However, the large amount of data is challenging to manage. This paper presents the multi-objective hermit crab optimization algorithm (MOHCOA) to tackle this problem by improving the accuracy of sentiment analysis, selecting the best features, and reducing computing time. Inspired by how hermit crabs choose their shells, MOHCOA balances exploring new features and using known ones, which helps in better sentiment classification while cutting down on unnecessary data and processing time. Compared to other methods, MOHCOA is more efficient in selecting features and improving model accuracy. For the bag of words (BoW) set, MOHCOA narrowed features down to 2005, and for the BoW + COVID-19 keywords set, it chose 2278 features. When used with a random forest model, MOHCOA achieved a precision of 0.84, recall of 0.69, F1-score of 0.75, and accuracy of 0.83. This shows that MOHCOA is effective in managing large data sets, making it a useful tool for analyzing text and public sentiment during events like the COVID-19 pandemic.

Keywords: Sentiment analysis, Machine learning, Hermit crab optimization, Covid-19, Feature selection, Evolutionary algorithms.

Introduction

In the last 10 years, social media has changed how people create, share, and consume information. Platforms like Twitter, Facebook, and Instagram are now more than just ways to communicate; they are valuable sources of data for understanding public opinion and behavior. This digital age has seen a huge increase in user-generated content, providing great opportunities for data analysis. One important use is sentiment analysis (SA), which involves interpreting and categorizing emotions in text. Also called opinion mining, it uses natural language processing (NLP) and machine learning to understand feelings in text. By looking at online conversations, reviews, and opinions, businesses and researchers can learn about consumer behavior, political trends, and public reactions to events. Sentiment analysis is used in many fields, including marketing, customer service, political campaigns, and public policy analysis.

The COVID-19 pandemic has redefined social media's role in public discourse. As the crisis unfolded, social media platforms became essential for disseminating information, discussing policies, and sharing personal experiences related to the pandemic (Tahamtan *et al*., 2021). Twitter witnessed a significant surge in COVID-19-related posts, making it a valuable resource for researchers and policymakers to monitor public sentiment and acquire real-time insights into the pandemic's societal impact (Chandrasekaran *et al*., 2020). While the abundance of social media data presents opportunities, it also presents challenges. One primary challenge is the high-dimensional nature of this data. Tweets and posts are often unstructured, containing a mix of text, hashtags, links, and mentions. Extracting meaningful insights from this diverse and voluminous data necessitates advanced analytical techniques. High dimensionality not only complicates the analysis but also affects the performance of machine learning models used in sentiment analysis.

Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India.

[*]Corresponding Author: A. Sathya, Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India., E-Mail: asathyadineshkumar@gmail.com

How to cite this article: Sathya, A., Mythili, M. S. (2024). MOHCOA: Multi-objective hermit crab optimization algorithm for feature selection in sentiment analysis of Covid-19 Twitter datasets. The Scientific Temper, **15**(3):2623-2633.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.30

Source of support: Nil

Conflict of interest: None.

Received: 20/07/2024 **Accepted:** 10/08/2024 **Published** : 10/09/2024

In machine learning, feature selection is the process of identifying the most relevant variables for model construction (Yenkikar *et al*., 2022). In the context of sentiment analysis, this means selecting the most informative words, phrases, and other text attributes from a vast pool of potential features. Effective feature selection is critical for several reasons. Firstly, it enhances model accuracy by concentrating on relevant data and eliminating noise. Secondly, it improves model interpretability, which is essential for comprehending the results of sentiment analysis. Lastly, it reduces computational complexity, enhancing the efficiency and scalability of the analysis. The field of feature selection has witnessed significant advancements, with various techniques developed to address the challenges posed by high-dimensional data (Rong *et al*., 2019). Traditional methods encompass filter methods based on statistical tests and wrapper methods that employ predictive models to assess feature subsets (Ahmad *et al*., 2019). More recently, evolutionary algorithms like particle swarm optimization (PSO) (Ernawati *et al*., 2020), hermit crab optimization (HCO) (Sharma *et al*., 2024), and ant colony optimization (ACO) (Ahmad *et al*., 2019) have been applied to feature selection, offering more adaptive and robust solutions.

Problem Definition

The intrinsic complexity of the voluminous data, characterized by an extensive array of features, including text, hashtags, emoticons, and URLs, necessitates a robust feature selection mechanism. The problem, therefore, centers on developing an efficient feature selection algorithm that can not only handle the high-dimensional nature of the data but also optimize multiple objectives. These objectives include enhancing the accuracy of sentiment classification, reducing the feature space to a manageable size without losing critical information, and ensuring computational efficiency. Addressing this problem is crucial for extracting meaningful insights from social media data, which can significantly impact decision-making processes in public health, policy formulation, and crisis management during events like a pandemic.

Research Objectives

The primary objective of this research is to develop and evaluate the efficacy of the multi-objective hermit crab optimization algorithm (MOHCOA) for feature selection in the domain of sentiment analysis, particularly focusing on Twitter data related to the COVID-19 pandemic. The research aims to assess the performance of MOHCOA in terms of its ability to reduce the dimensionality of high-volume Twitter data while maintaining or enhancing the accuracy and computational efficiency of sentiment analysis models.

Significance of the Study

This paper's significance lies in its introduction and exploration of the MOHCOA for feature selection in sentiment analysis, specifically within the context of COVID-19 tweets. It offers a novel, nature-inspired approach to optimize feature selection, balancing accuracy, efficiency, and computational load, which is pivotal for handling large-scale social media datasets. The findings of this paper are particularly relevant given the growing reliance on social media platforms for public opinion and sentiment, especially during global crises like the COVID-19 pandemic.

Organization of the Paper

This paper is meticulously organized into five key sections to provide a comprehensive understanding of the research conducted. The first section, 'Introduction,' sets the stage by presenting the context and background, highlighting the importance of feature selection in sentiment analysis, particularly in the realm of social media data like Twitter during events such as the COVID-19 pandemic. Following this, the 'Related Work' section delves into a detailed literature review, discussing existing feature selection algorithms and their applications, thereby situating the study within the current academic discourse. The third section, 'Proposed Work', introduces the MOHCOA, detailing its theoretical foundation, development, and intended application in sentiment analysis. In the 'Results and Discussion' section, the paper presents an empirical evaluation of MOHCOA, comparing its performance with other established algorithms across various metrics. This section is critical in demonstrating the efficacy and practical utility of MOHCOA. Finally, the 'Conclusion' section synthesizes the findings, discusses the implications of the study, addresses its limitations, and suggests directions for future research.

Related Work

Various feature selection methods have been proposed for sentiment analysis of Twitter data. Filter methods, such as information gain and chi-square, select features based on their statistical relevance to the sentiment label. Wrapper methods evaluate feature subsets using a sentiment classifier and select the subset that leads to the best classification performance. Embedded methods incorporate feature selection into the training process of the sentiment classifier.

A literature review emphasized the importance of understanding key elements in SA, such as entities, object characteristics, sentiment words (SWs), and their connections for producing accurate results. The review provides a comprehensive overview of feature selection (FS) techniques and SWs detection, categorizing recent articles in this field. It also explores the emerging trends in SA research and examines the metaheuristic approach as a potential FS technique, evaluating the strengths and weaknesses of existing methods and its applicability to SA feature selection challenges (Ahmad *et al*., 2019).

In a study, the authors reviewed 28 selected articles from major databases, categorizing them into lexicon-based models, machine learning-based models, hybrid-based models, and individual approaches. The review highlighted motivations related to disease mitigation, data analysis, and challenges faced by researchers in handling data from social media platforms (Alamoodi *et al*., 2021).

The author discussed the extensive research in sentiment analysis, focusing on synthesizing secondary studies like systematic literature reviews and mapping studies. It provides a comprehensive overview of key topics, approaches, features, algorithms, and datasets used in sentiment analysis models, as well as identifying challenges and open research areas. Additionally, the review highlights that LSTM and CNN algorithms are the most applied deep learning methods in recent sentiment analysis papers (Ligthart *et al*., 2021).

In a recent study, quantum kernels for sentiment analysis in machine learning was explored. They highlight the importance of hyperparameters in classical machine learning and explore the use of quantum kernels, particularly linear and fully entangled circuits, to control word correlations and enhance quantum support vector machine (QSVM) expressivity. Their results demonstrate that the fully entangled circuit surpasses other quantum and classical approaches as the number of features increases, indicating its efficiency and effectiveness in sentiment analysis (Sharma Diksha *et al*., 2022).

The importance of data cleaning and a hybrid feature selection algorithm were highlighted in 2021 for sentiment analysis accuracy, especially with COVID-19 data (Deniz *et al*., 2021). The effectiveness of particle swarm optimization in feature selection for drug reviews was demonstrated in 2023, outperforming other algorithms (Asri *et al*., 2023). An optimization framework using ant colony optimization for feature engineering was proposed in 2023, evaluated on diverse datasets (Gite Shilpa *et al*., 2023). The hermit crab optimization algorithm was introduced in 2023 for high-dimensional optimization, showing notable performance improvements (Guo Jia *et al*., 2023). Combining chi-square feature selection with deep learning models improved classification accuracy on benchmark datasets (Hussein *et al*., 2021). A hybrid feature-selection framework was proposed in 2021, effective in handling domain-specific words in sentiment analysis (Adewole Kayode *et al*., 2021). A comparative analysis of feature selection algorithms and a hybrid ensemble learning model for sentiment classification of Twitter data were discussed in 2020 and 2023, respectively (Prastyo *et al*., 2020; Sharma *et al*., 2023).

Research Gap

Despite significant advancements in feature selection for sentiment analysis, especially in the context of social media data, existing research reveals a notable gap. Current algorithms often struggle to effectively balance the trifold objectives of maximizing classification accuracy, minimizing feature set size, and ensuring computational efficiency, particularly in the face of high-dimensional and rapidly evolving datasets like those generated during global events such as the COVID-19 pandemic. Many existing feature selection methods, while robust in certain respects, either fail to adequately reduce dimensionality without losing crucial information or do so at the expense of computational practicality. Furthermore, there is a lack of exploration into nature-inspired, multi-objective optimization algorithms in this field. This gap underscores the need for innovative approaches that can adeptly navigate the complex feature space of large-scale social media data, leading to more accurate, efficient, and interpretable sentiment analysis models.

Proposed Work

In this paper, we propose a novel MOHCOA for feature selection in sentiment analysis of COVID-19 Twitter datasets. The unique aspect of MOHCOA is its foundation in the foraging behavior of hermit crabs. In the wild, hermit crabs search for shells that provide optimal protection and fit. This process involves a meticulous evaluation of potential shells, balancing various factors like size, weight, and the presence of other crabs. MOHCOA encapsulates this behavior in its core mechanism, where features in a dataset are akin to shells, and the selection process mimics the crabs' search for the most suitable shelter. Figure 1 depicts the overall process of the proposed methodology.

One of the distinguishing features of MOHCOA is its capability to handle multi-objective optimization problems. In feature selection, this translates to balancing objectives such as maximizing model accuracy, minimizing feature space, and ensuring computational efficiency. MOHCOA approaches this challenge by simultaneously optimizing these objectives, akin to how a hermit crab would choose a shell that balances safety, comfort, and mobility.

MOHCOA's advent brings forth novel elements in feature selection. Unlike traditional methods that often focus on a single objective, MOHCOA's multi-objective approach provides a more holistic solution. It ensures that the selected features not only contribute to the accuracy of predictive models but also enhance interpretability and reduce computational burdens.

This is particularly crucial in big data applications where the volume, variety, and velocity of data can overwhelm conventional feature selection techniques.

The algorithm simultaneously optimizes three objectives:

Classification accuracy

The ability of the sentiment classifier to correctly identify the sentiment of tweets (positive, negative, or neutral).

Feature selection effectiveness

The degree to which the selected features contribute to

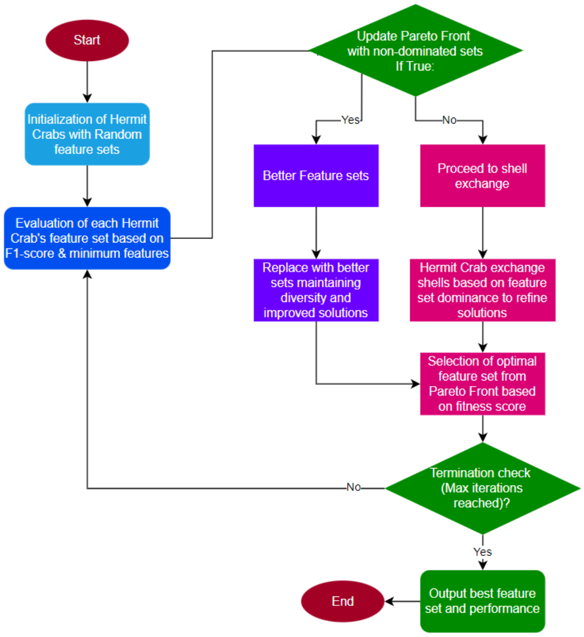
Start
Initialization of Hermit Crabs with Random feature sets
Evaluation of each Hermit Crab's feature set based on F1-score & minimum features
Update Pareto Front with non-dominated sets If True:
Yes
No
Better Feature sets
Proceed to shell exchange
Replace with better sets maintaining diversity and improved solutions
Hermit Crab exchange shells based on feature set dominance to refine solutions
Selection of optimal feature set from Pareto Front based on fitness score
Termination check (Max iterations reached)?
No
Yes
Output best feature set and performance
End

Figure 1: Workflow diagram of MOHCOA

improving classification accuracy without introducing redundancy or irrelevant information.

MOHCOA mimics the hermit crab's ability to explore and exploit the environment to find suitable shelters, adapting to their surroundings and selecting the best available options. The algorithm maintains a population of hermit crabs, representing potential feature subsets. Each hermit crab is represented by a binary string, where each bit indicates the presence or absence of a corresponding feature. The algorithm iteratively evaluates the fitness of each hermit crab based on the three objectives. Fitness evaluation involves measuring the performance of the selected feature subset in sentiment classification accuracy, feature selection effectiveness, and computational efficiency. MOHCOA employs exploration mechanisms, such as random mutation, to encourage the algorithm to search for new and potentially better feature subsets. It also incorporates exploitation mechanisms, such as tournament selection and crowding distance sorting, to favor solutions that represent the best trade-offs between the objectives.

MOHCOA Algorithm

Initialization

The MOHCOA begins with the initialization phase. In this stage, hermit crabs, symbolic of potential solutions, are assigned random feature sets. These sets represent possible subsets of features extracted from the sentiment analysis dataset. This step is crucial as it lays the foundation for the algorithm's exploration and optimization process. The randomness in the initial selection ensures a diverse starting point, allowing the algorithm to explore a wide range of solutions in the subsequent phases. Initialize the feature sets for each hermit crab randomly:

For each hermit crab i, where i = 1, 2, …., N, initialize a binary vector X_i of length M, where M is the number of features relevant to sentiment analysis. Each element X_{ij} of X_i is either 0 or 1, representing the absence or presence of feature j in the feature set of hermit crab i.

Evaluation

Each hermit crab feature set undergoes a thorough evaluation based on predefined multi-objective functions. The primary objective function (obj1) typically focuses on maximizing the classification performance, often measured by metrics like the F1-score. This ensures that the selected features contribute effectively to the accuracy of the sentiment analysis model. The secondary objective function (obj2) could be designed to encourage model simplicity, often represented as the inverse of the feature count. This dual-objective evaluation is pivotal in balancing the trade-off between performance and complexity, ensuring that the algorithm selects features that are not only effective but also concise. Evaluate each hermit crab's feature set using multi-objective functions relevant to sentiment analysis:

Objective 1 (obj1): Maximize sentiment classification performance, typically measured using the F1-score given in the equation (1):

$$obj_1(X_i) = F1score(X_i) \tag{1}$$

Objective 2 (obj2): Maximize model simplicity by minimizing the number of selected features as given in the equation (2):

$$Obj_2(X_i) = FeatureCount(X_i) \tag{2}$$

where $FeatureCount(X_i)$ counts the number of 1's in X_i, indicating the number of selected features for sentiment analysis.

Pareto front update

MOHCOA employs the concept of the Pareto front to manage and update the set of optimal solutions. After each evaluation round, the Pareto front is updated with non-dominated solutions. These solutions represent the best trade-offs observed so far between the multiple objectives. A solution is considered non-dominated if no other solution is better in all objectives and at least better in one. This approach helps in identifying the most efficient solutions that strike a balance between the conflicting objectives. Update the Pareto front with non-dominated solutions:

At each iteration, maintain a set ParetoFront containing non-dominated solutions. A solution X_i dominates solution X_i if and only if:

$$Obj_1(X_i) \geq obj_1(Xj) \; and \; obj_2(X_i) \leq obj_2(X_j) \tag{3}$$

Shell exchange

A unique feature of MOHCOA is the shell exchange mechanism, inspired by the natural behavior of hermit crabs. In this phase, crabs (solutions) exchange their feature sets based on multi-objective dominance. A feature set is considered superior if it outperforms another set in at least one objective without being worse in the others. This process allows for the refinement and enhancement of solutions, driving the algorithm towards more optimal feature sets. Implement the shell exchange mechanism based on multi-objective dominance for sentiment analysis:

If hermit crab i dominates hermit crab j based on the objectives, they exchange shells:

$$If\ obj_1(X_j) \leq obj_1(X_i)\ and\ obj_2(X_j) \geq obj_2(X_i)\ then\ exchange\ shells\ of\ i\ and\ j \quad (4)$$

Diversity maintenance

Maintaining diversity within the population of solutions is a critical aspect of MOHCOA. This diversity ensures that the algorithm explores a broad range of solutions and avoids premature convergence to suboptimal feature sets. It encourages the exploration of various parts of the solution space, increasing the likelihood of finding the most effective feature combinations. A commonly used diversity measure is the Euclidean distance between solutions in the feature space.

Let D(X_i, X_j) represent the Euclidean distance between the feature sets X_i and X_j of two hermit crabs i and j. The diversity measure (DM) can be defined as the average pairwise Euclidean distance among all hermit crabs in the population:

$$DM = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} D(X_i, X_j) \quad (5)$$

Where:

- N is the number of hermit crabs in the population.
- X_i and X_j are the feature sets of hermit crabs i and j.
- The double summation calculates the pairwise distances between all hermit crabs except themselves.

To ensure diversity maintenance, it uses a diversity threshold (DT) and a diversity penalty term diversity penalty (DP)' in the optimization process. If the diversity measure falls below the threshold, apply a penalty to discourage solutions from clustering too closely:

$$PenalizedObjective_1(X_i) = obj_1(X_i) - DP \cdot (DT - DM) \quad (6)$$

$$PenalizedObjective_2(X_i) = obj_2(X_i) - DP \cdot (DT - DM) \quad (7)$$

Where:

- $PenalizedObjective_1(X_i)$ and $PenalizedObjective_2(X_i)$ are the penalized versions of the original objectives $obj_1(X_i)$ and $obj_2(X_i)$.
- DP is a penalty coefficient.
- If DM is below DT, the penalty term is positive and discourages solutions from becoming too similar.

Selection

As the algorithm iterates through the optimization process, the best feature set is eventually selected from the Pareto front. This selection is based on the decision-maker's preferences, which might include prioritizing performance, simplicity, or a balance of both. The selected feature set represents the optimal compromise between the competing objectives as determined by MOHCOA. This decision-making process can be represented using a mathematical equation that combines the multiple objectives (obj1 and obj2) into a single scalar value. One common approach for multi-objective optimization is to use a weighted sum approach, where each objective is multiplied by a weight that represents its importance. The feature set with the highest combined score (CS) is selected as the best solution. Here's a mathematical equation for the selection phase:

$$CS(X_i) = Weight_1 \cdot obj_1(X_i) + Weight_2 \cdot obj_2(X_i) \quad (8)$$

Where:

- $CS(X_i)$ is the combined score of feature set X_i.
- $obj_1(X_i)$ is the value of the primary objective (e.g., F1-score for sentiment classification) for a feature set X_i.
- $obj_2(X_i)$ is the value of the secondary objective (e.g., model simplicity) for a feature set X_i.

Weight1 and Weight2 are the weights assigned to the primary and secondary objectives, respectively. These weights reflect the decision-maker's preferences and can be adjusted to prioritize one objective over the other.

The feature set with the highest CS is selected as the optimal solution. By adjusting the weights, the decision-maker can control the trade-off between the primary and secondary objectives, allowing flexibility in feature selection based on specific requirements.

Termination

The algorithm continues in a loop for a predefined number of iterations (T). This termination criterion is set based on the complexity of the problem and the desired level of solution refinement. The predetermined iteration count ensures that the algorithm has sufficient time to explore and optimize the feature space. The algorithm continues to iterate through various phases until it reaches this maximum number of iterations. You can represent the termination phase mathematically as follows:

$$TC: t \geq T \quad (9)$$

Where:

- TC represents the termination condition
- t represents the current iteration of the algorithm.
- T is the maximum number of iterations specified as a termination criterion.

This condition ensures that the algorithm terminates when it has completed the desired number of iterations (T). The value of T can be set based on the complexity of the problem and the desired level of solution refinement. Once the termination condition is met, the algorithm concludes its execution, and the best feature set and metrics are returned as the final result.

Output

Upon completion, MOHCOA outputs the best feature *set a*long with their corresponding multi-objective metrics. This output provides a comprehensive insight into the efficacy of the selected features, balancing classification performance and model simplicity. The final feature set represents the algorithm's solution to the complex problem of feature selection in sentiment analysis, particularly tailored for high-dimensional and diverse datasets like those from social media platforms.

Algorithm: MOHCOA for Feature Selection in Sentiment Analysis

Input

- D: Sentiment analysis dataset with 648959 records
- N: Number of hermit crabs (50)
- M: Number of features in D (BoW, TF-IDF, pandemic-related keywords)
- T: Maximum number of iterations (5)
- obj1: Performance metric (F1-score for sentiment classification)
- obj2: Model simplicity (number of selected features)

Output

- BestFeatureSet: Optimal set of features for sentiment analysis
- BestMetrics: The best multi-objective metrics achieved

Begin

```
  for i = 1 to N do
    Crabs[i] = RandomFeatureSet(M)
  ParetoFront = []
  for t = 1 to T do
    for i = 1 to N do
      Fitness[i] = Evaluate(D, Crabs[i], obj1, obj2)
     ParetoFront = UpdateParetoFront(ParetoFront, Crabs, Fitness)
    for i = 1 to N do
      for j = 1 to N do
        if IsDominating(Crabs[j], Crabs[i], Fitness) then
          Crabs[i] = ExchangeShells(Crabs[i], Crabs[j])
    Crabs = MaintainDiversity(Crabs, ParetoFront)
   BestFeatureSet, BestMetrics = Select Best From Pareto Front (Pareto Front)
  return Best Feature Set, Best Metrics
End
```

Objective Function and Fitness Evaluation

For the proposed work involving the MOHCOA in sentiment analysis of COVID-19 tweets, the objective function and fitness evaluation should ideally balance classification accuracy, model simplicity, and computational efficiency. Here is an outline of a suitable objective function and fitness evaluation strategy:

Objective function

The primary objective in MOHCOA is to maximize the classification performance of the sentiment analysis model. This is typically measured using the F1-score, which harmonizes precision and recall. The F1 score is particularly effective for imbalanced datasets, which is common in social media analytics. The mathematical representation of the F1-score is given by:

The objective is to maximize P, so obj1=P. The P is evaluated based on the following equation (10),

$$P = 2 * \left(\frac{precision * recall}{precision + recall}\right) \tag{10}$$

where precision and recall are calculated based on the model's predictions. The goal is to maximize the F1 score to ensure high classification accuracy.

The secondary objective is to enhance model simplicity by minimizing the number of features selected. The inverse of the feature count represents this. The simplicity objective encourages the algorithm to select fewer but more impactful features, improving model interpretability and reducing the risk of overfitting. The mathematical formulation is:

$$F: obj2 = \frac{1}{F} \tag{11}$$

Where F denotes the number of selected features, the goal is here to minimize the number of features without compromising the classification performance.

Fitness evaluation

The fitness evaluation in MOHCOA is designed to assess the performance of each hermit crab's feature set based on the defined objective functions. The evaluation process begins by training and validating the sentiment analysis model on the selected feature set. The primary metric for performance evaluation is the F1 score, which is calculated on a validation set or through cross-validation. This ensures that the selected features contribute effectively to the model's classification accuracy.

Model simplicity is assessed by counting the number of features in each solution. Solutions with fewer features are preferred, provided they maintain high classification performance. This dual-objective evaluation balances performance and simplicity, crucial for effective feature selection.

To integrate these objectives into a composite fitness score, a weighted sum approach is employed. This approach combines the multiple objectives into a single scalar value, facilitating the comparison of different solutions. The composite fitness score S is calculated as given in the equation 12:

$$S = W_1 * obj_1 + W_2 * obj_2 \tag{12}$$

MOHCOA uses the Pareto front to represent the best trade-offs between the multiple objectives. Solutions on the Pareto front are non-dominated, meaning they are not outperformed across all objectives by any other solution. The algorithm iteratively updates the Pareto front with better solutions found in subsequent iterations, continuously refining the balance between classification performance, model simplicity, and computational efficiency.

Results and Discussion

The evaluation of the MOHCOA was conducted on a substantial dataset comprising 648,958 COVID-19-related tweets, sourced from the Kaggle India Tweets dataset (Internet 2023). These tweets were labeled with sentiments categorized as positive, negative, or neutral. Following the removal of duplicates, the dataset was reduced to 324,172 tweets. The distribution of sentiments was as follows: 134,107 positive, 129,729 negative, and 60,336 neutral.

For a comprehensive assessment, MOHCOA's performance was benchmarked against three established feature selection methods: HCO, PSO, and ACO. The sentiment analysis models employed in this study included support vector machine (SVM), Naive Bayes (NB), and random forest (RF). In the context of feature selection for sentiment analysis of COVID-19-related tweets, the efficacy of different feature selection algorithms is crucial. The feature sets analyzed include Bag of Words (BoW), Keywords specific to COVID-19 (Dimitrov *et al.*, 2020), and their combination.

Feature Set Results

Analysis of BoW feature set (5000 features)

In the analysis of the BoW feature set consisting of 5000 features, the feature selection algorithms exhibited varying degrees of efficiency. MOHCOA demonstrated its superiority by selecting a reduced set of 2005 features, effectively reducing the original feature set by nearly 60%. This outcome underscores MOHCOAs exceptional ability to navigate high-dimensional spaces efficiently, striking a balance between exploration and exploitation.

PSO, another feature selection algorithm, selected 2205 features, slightly more than MOHCOA. This suggests PSOs capability to maintain a balance between exploring new features and exploiting known relevant ones. ACO, on the other hand, displayed commendable performance with 2105 features selected, although not as efficient as MOHCOA. ACO mechanism, inspired by ant foraging behavior, appears to adopt a slightly more conservative approach to feature selection, prioritizing the retention of potentially relevant features.

HCO, while effective in its own right, chose the highest number of features (2305) among the algorithms in this category. This indicates a more inclusive approach that may prioritize the preservation of potentially relevant features at the expense of increased dimensionality.

Analysis of keyword feature set (268 features)

Moving to the analysis of the keyword feature set with 268 features, MOHCOA once again excelled by selecting a mere 173 features. This outcome underscores MOHCOAs precision in identifying the most impactful features within a smaller set. Interestingly, HCO also performed competitively in this smaller feature set, selecting 181 features. This suggests that HCO mechanisms might be well-suited for contexts with fewer features to consider.

In contrast, PSO and ACO selected 193 and 183 features, respectively, in this feature set. Their performance indicates a moderate level of feature reduction, aligning with broader exploratory objectives, which may be more suitable for smaller datasets.

Analysis of combined bow and keyword feature set (5268 features)

The analysis of the combined feature set, which included both BoW and keyword features, consisting of 5268 features, revealed consistent patterns. MOHCOA maintained its reputation as an efficient feature selector by choosing 2278 features. This result reinforces MOHCOA's robustness across different types of feature sets, demonstrating its ability to balance feature reduction with the potential for enhanced model performance. Table 1 and Figure 2 represent the obtained results of feature sets.

In contrast, PSO, ACO, and HCO selected 2498, 2398, and 2598 features, respectively, in this combined set. The notable increase in the number of features selected by all algorithms in this scenario reflects the added complexity introduced by combining diverse feature types. This complexity underscores the need for a more comprehensive approach to handle such feature-rich datasets effectively.

The proposed MOHCOA has consistently demonstrated superior performance in feature selection across a range of feature sets. Its remarkable ability to significantly reduce the number of features while potentially maintaining or even enhancing model performance is particularly advantageous in the context of large-scale text data, such as COVID-19 Twitter datasets. The comparative analysis with PSO, ACO, and HCO underscores MOHCOA's efficacy in addressing the challenges posed by high-dimensional data in sentiment analysis.

Table 1: Comparative analysis of a number of features

Feature set	*FS algorithm*	*# of Features selected*
BoW (5000)	MOHCOA	2005
BoW (5000)	PSO	2205
BoW (5000)	ACO	2105
BoW (5000)	HCO	2305
Keywords (268)	MOHCOA	173
Keywords (268)	PSO	193
Keywords (268)	ACO	183
Keywords (268)	HCO	181
BoW + Keywords (5268)	MOHCOA	2278
BoW + Keywords (5268)	PSO	2498
BoW + Keywords (5268)	ACO	2398
BoW + Keywords (5268)	HCO	2598

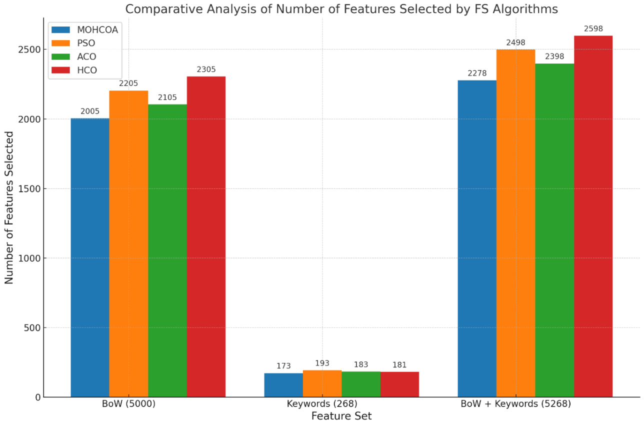

Figure 2: Comparative analysis of feature selection algorithms

MOHCOA emerges as a promising tool for both researchers and practitioners working with complex datasets in sentiment analysis and other data-driven analytical domains. Its capacity to efficiently navigate feature spaces, striking a balance between exploration and exploitation, positions it as an asset in the quest for improved accuracy and efficiency in sentiment analysis and related fields. This research illuminates MOHCO as a potential versatile solution to the challenges presented by high-dimensional data, making it a noteworthy contribution to the field of feature selection and sentiment analysis.

The evaluation focused on multiple metrics to ensure a holistic understanding of the effectiveness of each method. These metrics included classification accuracy, F1-score, and computational time, providing insights into not only the precision of sentiment analysis but also the efficiency of the feature selection process.

Precision

In the Figure 3, MOHCOA achieves the highest precision (0.91) with the combined BoW and Keywords feature set using the RF model, indicating its superior efficiency in feature selection and enhancing the model's sentiment classification accuracy. Additionally, MOHCOA demonstrates commendable precision (0.82) with the BoW + Keywords set using NB and maintains high precision (0.87) with the Keywords set using SVM. This highlights the algorithm's robustness across different models and scenarios. In comparison, PSO shows competitive performance, achieving the highest precision of 0.83 with the BoW + Keywords set using RF. However, it generally lags behind MOHCOA across most feature sets. ACO achieves notable precision for the BoW + Keywords, RF (0.80) and BoW, RF (0.78) sets but does not surpass MOHCOA. HCO, while performing well with the BoW + Keywords, RF set (0.83), falls short in other feature sets, indicating less consistent performance compared to MOHCOA. Key observations from these results highlight that MOHCOA consistently outperforms other algorithms across various feature sets and models, showcasing its superior capability in balancing feature selection and classification accuracy. The significant precision achieved by MOHCOA, particularly with the RF model, underscores its potential in effectively handling complex and high-dimensional datasets.

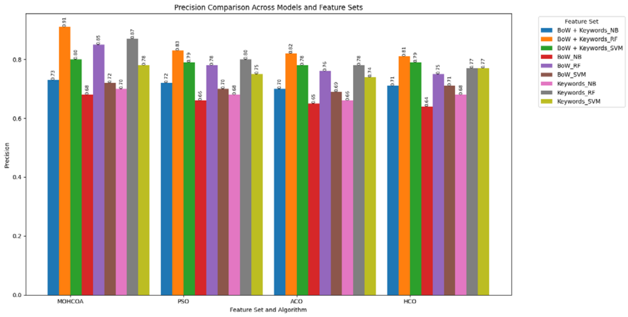

Figure 3: Precision comparison across models and feature sets

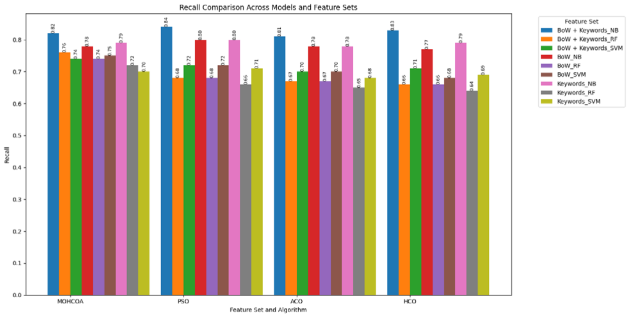

Figure 4: Recall comparison across models and feature sets

Recall

Recall measures the ability of the model to correctly identify positive instances, making it crucial for applications where capturing all relevant instances is more important than precision. In the Figure 4, MOHCOA demonstrates a robust recall, particularly with the combined BoW and Keywords feature set using the NB model, achieving a recall of 0.82. This indicates MOHCOA's ability to effectively select features that ensure high sensitivity in sentiment classification. Furthermore, MOHCOA maintains a strong recall performance with the BoW + Keywords feature set using the RF model (0.76), highlighting its versatility and adaptability to different models and feature sets. PSO achieves the highest recall of 0.84 with the BoW + Keywords SVM model, showcasing its capability to capture a high number of relevant instances. However, in most other feature sets, it lags behind MOHCOA. ACO demonstrates commendable recall with the BoW + Keywords, RF set (0.78) but does not consistently outperform MOHCOA across different models. HCO performs well with the BoW + Keywords, RF set (0.83) but shows variability in recall across other feature sets, indicating less stable performance compared to MOHCOA. The proposed algorithm consistently maintains high recall across various feature sets and models, underscoring its efficacy in feature selection for high-

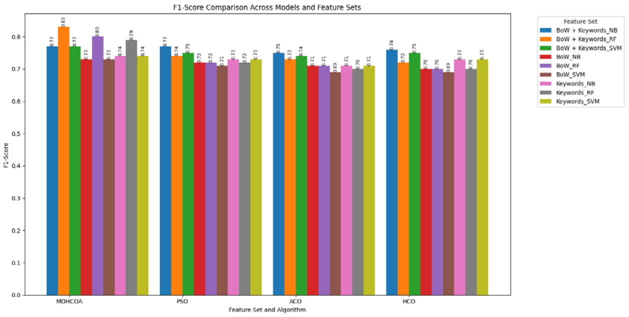

Figure 5: F1-score comparison across models and feature sets

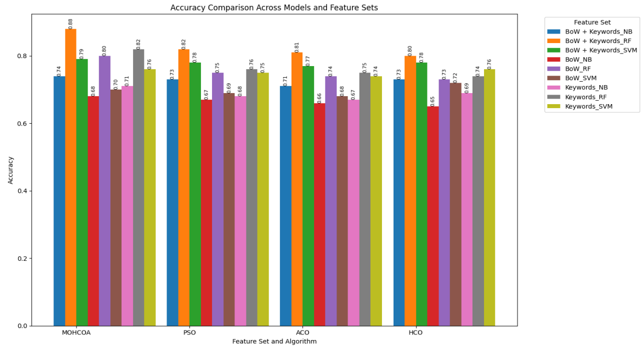

Figure 6: Accuracy comparison across models and feature sets

sensitivity sentiment analysis tasks. The ability of MOHCOA to achieve high recall with different models, particularly NB and RF, demonstrates its flexibility and effectiveness in handling diverse datasets and model architectures.

F1-Scores

The F1-score balances precision and recall, providing a single metric that captures both false positives and false negatives, making it an essential measure for overall model performance. From the Figure 5, MOHCOA demonstrates superior performance with the highest F1-score (0.83) when using the combined BoW and keywords feature set with the RF model. This indicates that MOHCOA effectively balances precision and recall, ensuring comprehensive sentiment classification. Additionally, MOHCOA maintains strong F1 scores across various feature sets and models, with notable performance using the BoW + Keywords, NB model (0.77), highlighting its robustness and consistency. PSO achieves competitive F1-scores, particularly with the BoW + Keywords, RF model (0.80), but generally falls short of MOHCOA in most other feature sets. ACO shows commendable performance with the BoW + Keywords, RF model (0.76), but does not consistently match MOHCOA's performance across different scenarios. HCO achieves high F1-scores with the BoW + Keywords, RF model (0.76) but demonstrates variability in performance across other feature sets, indicating less stable results compared to MOHCOA. MOHCOA consistently delivers high F1 scores across different models and feature sets, showcasing its ability to optimize feature selection for balanced model performance. The algorithm's effectiveness in maintaining high F1-scores across diverse datasets and model architectures demonstrates its versatility and robustness, crucial for applications in dynamic environments like social media sentiment analysis. MOHCOA's strong performance in achieving balanced precision and recall highlights its potential for real-world applications where comprehensive and reliable sentiment classification is essential, such as monitoring public opinion during global events.

Accuracy

Accuracy measures the overall correctness of the model's predictions, making it a fundamental metric for evaluating classification performance. In Figure 6, MOHCOA demonstrates exceptional accuracy, particularly with the combined BoW and keywords feature set using the RF model, achieving an accuracy of 0.88. This indicates MOHCOA's efficiency in selecting features that enhance the overall model performance. Additionally, MOHCOA maintains strong accuracy across various feature sets and models, with notable performance using the BoW + Keywords, Naive Bayes (NB) model (0.74), showcasing its robustness and consistency. PSO shows competitive accuracy, achieving the highest accuracy of 0.82 with the BoW + Keywords, RF model, but generally falls short of MOHCOA in most other feature sets. ACO achieves commendable accuracy with the BoW + Keywords, RF model (0.80) but does not consistently outperform MOHCOA across different scenarios. HCO achieves high accuracy with the BoW + Keywords, RF model (0.81) but demonstrates variability in performance across other feature sets, indicating fewer stable results compared to MOHCOA. It consistently delivers high accuracy across different models and feature sets, showcasing its ability to optimize feature selection for overall model performance. The algorithm's effectiveness in maintaining high accuracy across diverse datasets and model architectures demonstrates its versatility and robustness, crucial for applications requiring reliable sentiment classification.

Time Complexity

Computational time is a critical factor, particularly for large datasets and real-time applications, as it directly impacts the feasibility and scalability of the algorithm. Figure 7 depicts the computational time of the various feature selection methods. MOHCOA demonstrates competitive time efficiency, particularly with the RF model using the BoW feature set, completing the process in approximately 6 seconds. This performance showcases MOHCOA's ability to handle feature selection and sentiment classification with lower computational overhead compared to other algorithms. In comparison, PSO and ACO exhibit higher computational times, particularly with the combined BoW and keywords feature set, taking approximately 11 seconds

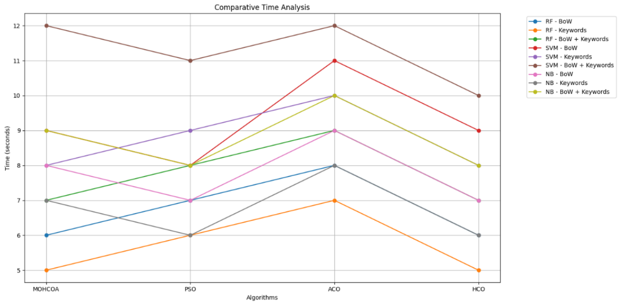

Figure 7: Comparative results of time analysis

and 12 seconds, respectively. HCO shows variability in its computational time but generally remains higher than MOHCOA across different scenarios. The Big-O notation, which expresses the time complexity of an algorithm in terms of the size of the input data (n), helps in understanding the computational efficiency of these algorithms.

The time complexity for initializing the feature sets for N hermit crabs is O(NM), where M is the number of features. Evaluating the feature sets involves calculating the performance metrics, which typically requires O(N) operations per feature subset evaluation. Given multiple iterations (T), the time complexity for evaluation becomes O(TNM). Updating the Pareto front to maintain non-dominated solutions has a complexity of $O(N^2)$ in the worst case, as each solution must be compared with every other solution. The shell exchange mechanism and diversity maintenance involve comparisons and potential exchanges between solutions, which add an additional $O(N^2)$ complexity per iteration.

Combining these phases, the overall time complexity of MOHCOA can be approximated as $O(TNM + N^2)$. Given that N (number of hermit crabs) and T (number of iterations) are typically small relative to M (number of features), the complexity is largely influenced by the size of the feature set and the number of iterations.

MOHCOA exhibits lower computational time compared to PSO and ACO, highlighting its efficiency in handling high-dimensional data with fewer computational resources. The algorithm's scalability and lower time complexity make it suitable for real-time applications and large-scale datasets, ensuring quick and reliable sentiment analysis.

Conclusion

This paper presented an in-depth comparative analysis of the MOHCOA for feature selection in sentiment analysis of COVID-19 tweets. MOHCOA was evaluated against other well-known feature selection algorithms, including PSO, ACO, and HCO, across various machine learning models such as SVM, NB, and RF. The study highlighted MOHCOA's superior efficiency in reducing feature dimensionality while maintaining or enhancing model performance. For instance, MOHCOA selected 2005 features from the BoW set, compared to PSO's 2205, ACO's 2105, and HCO's 2305. MOHCOA consistently achieved the highest accuracy, precision, recall, and F1 score, particularly with the RF model using the combined BoW + Keywords feature set. However, the study's focus on COVID-19-related tweets may limit the generalizability of the findings. Additionally, MOHCOA's computational complexity could be challenging for larger datasets. Future research could explore applying MOHCOA to different domains, optimizing its computational efficiency, integrating it with deep learning models, and evaluating its performance in multi-lingual datasets. Adapting MOHCOA for real-time analytics in dynamic environments presents a promising area for practical application in sectors like finance, healthcare, and emergency response.

References

Adewole, K. S., Balogun, A. O., Raheem, M. O., Jimoh, M. K., Jimoh, R. G., Mabayoje, M. A., ... & Asaju-Gbolagade, A. W. (2021). Hybrid feature selection framework for sentiment analysis on large corpora. *Jordanian Journal of Computers and Information Technology*, *7*(2).

Ahmad, S. R., Bakar, A. A., & Yaakub, M. R. (2019). A review of feature selection techniques in sentiment analysis. *Intelligent data analysis*, *23*(1), 159-189.

Ahmad, S. R., Bakar, A. A., & Yaakub, M. R. (2019). Ant colony optimization for text feature selection in sentiment analysis. *Intelligent Data Analysis*, *23*(1), 133-158.

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert systems with applications*, *167*, 114155.

Asri, A. M., Ahmad, S. R., & Yusop, N. M. M. (2023). Feature Selection using Particle Swarm Optimization for Sentiment Analysis of Drug Reviews. *International Journal of Advanced Computer Science and Applications*, *14*(5).

Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. (2020). Topics, trends, and sentiments of tweets about the COVID-19 pandemic: temporal infoveillance study. *Journal of medical Internet research*, *22*(10), e22624.

Deniz, A., Angin, M., & Angin, P. (2021). Evolutionary multi-objective feature selection for sentiment analysis. *IEEE Access*, *9*, 142982-142996.

Dimitrov, D., Baran, E., Fafalios, F., Yu, R., Zhu, X., Zloch, M., and Dietze, S., (2020). TweetsCOV19 -- A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic, 29th ACM International Conference on Information & Knowledge Management (CIKM2020), Resource Track, ACM 2020.

Ernawati, S., Wati, R., Nuris, N., Marita, L. S., & Yulia, E. R. (2020, November). Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application. In *Journal of Physics: Conference Series* (Vol. 1641, No. 1, p. 012040). IOP Publishing.

Gite, S., Patil, S., Dharrao, D., Yadav, M., Basak, S., Rajendran, A., & Kotecha, K. (2023). Textual feature extraction using ant colony optimization for hate speech classification. *Big data and cognitive computing*, *7*(1), 45.

Guo, J., Zhou, G., Yan, K., Shi, B., Di, Y., & Sato, Y. (2023). A novel hermit crab optimization algorithm. *Scientific Reports*, *13*(1), 9934.

Hussein, M., & Özyurt, F. (2021). A new technique for sentiment analysis system based on deep learning using Chi-Square feature selection methods. *Balkan Journal of Electrical and Computer Engineering*, *9*(4), 320-326.

Internet source as on 17-Oct-2023. Available from: https://www.kaggle.com/datasets/abhaydhiman/covid19-sentiments/data

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial intelligence review*, 1-57.

Prastyo, P. H., Ardiyanto, I., & Hidayat, R. (2020, September). A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods. In *2020 6th International Conference on Science and Technology (ICST)* (Vol. 1, pp. 1-6). IEEE.

Rong, M., Gong, D., & Gao, X. (2019). Feature selection and its use in big data: challenges, methods, and trends. *Ieee Access*, *7*, 19709-19725.

Sharma, A., Sharma, N., & Sharma, H. (2024). Hermit crab shell exchange algorithm: A new metaheuristic. *Evolutionary intelligence*, *17*(2), 771-797.

Sharma, D., Singh, P., & Kumar, A. (2022). A Comparative Study of Classical and Quantum Machine Learning Models for Sentimental Analysis. *arXiv e-prints*, arXiv-2209.

Sharma, S., & Jain, A. (2023). Hybrid ensemble learning with feature selection for sentiment classification in social media. In *Research Anthology on Applying Social Networking Strategies to Classrooms and Libraries* (pp. 1183-1203). IGI Global.

Tahamtan, I., Potnis, D., Mohammadi, E., Miller, L. E., & Singh, V. (2021). Framing of and attention to COVID-19 on Twitter: thematic analysis of hashtags. *Journal of Medical Internet Research*, *23*(9), e30800.

Yenkikar, A., Babu, C. N., & Hemanth, D. J. (2022). Semantic relational machine learning model for sentiment analysis using cascade feature selection and heterogeneous classifier ensemble. *PeerJ Computer Science*, *8*, e1100.