RESEARCH ARTICLE OPEN ACCESS

An Efficient Comparative Study of Genetic Algorithm and Particle Swarm Optimization for Feature Selection in Sentiment Analysis Using MOOC Dataset

S. Daisy Fatima Mary
Research Scholar,
Department of Computer and Information Science,
Annamalai University, Annamalai Nagar
Dr. G. Mageswary
Assistant Professor,
Department of Computer Science,
PSPT MGR Government Arts and Science College,
Sirkali, Puthur
India

ABSTRACT

For delivering a large-scale education and also generating huge volumes of learner feedback and reviews in the form of textual data, nowadays, the Massive Open Online Courses (MOOCs) have emerged a leading platform. The proposed paper presents a comparative study of Genetic Algorithm and Particle Swarm Optimization for effective feature selection wrapper based feature selection to perform sentiment Analysis for MOOC datasets. The algorithms are mainly applied to reduce the sparse and high-dimensional feature space using machine learning classifiers such as Logistic Regression, SVM, Naïve Bayes, Random Forest, and Passive Aggressive Classifier. Both algorithms were implemented using binary encodings and a fitness function combining stratified 5-fold F1-score with a small subset-size penalty. The results show that optimization-based feature selection substantially improves accuracy, precision, recall, and F1-score compared to models without feature selection. GA proves in strong exploration ability and often selects a larger feature subset, which yielding to higher accuracy for certain classifiers. The experimental results show that PSO is faster in converges and produces compressed feature subsets which holds best classification performance also GA sometimes achieves higher F1 score values in the evaluations. The statistical test confirms that the PSO repeatedly selects less features of selected subsets are significantly smaller than the ($p < 0.05$), while differences in F1 are not statistically significant ($\alpha = 0.05$). Overall, the comparative study concludes that GA provides better exploration while PSO offers faster convergence and computational efficiency making PSO more suitable for large-scale sentiment-analysis applications.

Keywords: Feature Selection, Genetic Algorithm, Particle Swarm Optimization, Sentiment Analysis, MOOC, Wrapper Method, Metaheuristic Optimization.

I. INTRODUCTION

The Massive Open Online Courses have been transformed the higher education by providing open access, scalability, and flexibility for learners worldwide. In the Sentiment Analysis the feature selection for high dimensional datasets such as MOOCs plays an important role in text-based

information. Huge amounts of textual data in the form of forum posts, feedback, and reviews are generated by Massive Open Online Courses. By performing sentiment analysis using Mooc course dataset it gives many benefits to the instructors and as well as to the educational institute such as used to measure student availability during the course completion, helps to identifies dissatisfaction or problems in the learning courses, and also notifies the continuous improvements in course content and teaching suggested by (Deng & Lai, 2023) [1]. The MOOC platform also creates new challenges such as very large scale dimensionality, learner's feedback, continuous monitoring and evaluating satisfaction assessment. These problem in high dimensionality will increase the computational cost, creates irrelevant or noisy features, and may also reduce classifier generalization. Traditional machine learning classifiers finds difficulties in handling the textual data with very high dimensionality to perform sentiment analysis. The student's platforms generally produce thousands and thousands of feedback regarding the course, instructor, institution and assessment all these reviews are not taken directly into the processing of sentiment analysis. Before the processing it need to perform perprocessing such as tokenization, TF-IDF or embedding conversion also produces many are redundant, irrelevant or noisy data it has to be further processed like cleaning and transformation. The paper (Chandrashekar et al., 2016) [2], suggests that the high dimensionality with irrelevant attributes may degrade the classification prediction and also sometimes the performance and increase computational cost is really the curse.

By facing these challenges in the computation process of the sentiment analysis, this paper suggests to handle metaheuristic algorithms such as the Genetic Algorithm and Particle Swarm Optimization to applied for feature selection and hyper parameter tuning in the analysis to perform text classification. Feature selection through GA and PSO helps to reduce the feature-space dimensionality, noise elimination, and optimize model parameters, which consequently improves the accuracy and efficiency in sentiment classification (Rafdi et al., 2021) [3]. The author (Srivastava et al., 2023) [4], proves the sentiment classification have been done through PSO and GA algorithm by using social media text but the results shows that the individual performance of both the algorithms is comparatively low and the hybrid models by combining GA and PSO increases in accuracy.

By selecting a subset of detailed features and discarding irrelevant data also reduces dataset dimensionality is done by the Feature selection. The Genetic Algorithm and Particle Swarm Optimization of Metaheuristic algorithms are emerged to perform effective feature selection using wrapper based methods. Genetic Algorithm generally computes natural feature selection and explore global feature interactions by genetic recombination, whereas PSO models processing techniques such as swarm intelligence, updating candidate solutions based on individual and global experience. This proposed work performs an effective comparative study of GA and PSO using to MOOC dataset to perform sentiment analysis and produces the high accuracy, F1-score, subset size, convergence behaviour, and computational cost for selecting certain feature selection methods.

Understanding the MOOC learner experiences through student's reviews and the feature selection methods are applied to perform sentiment classification and prediction. The analysing with the prior work there is no systematic comparative analysis was done by using these two algorithms specifically to perform sentiment classification within the context of MOOC dataset but it has found that GA based or PSO based feature selection done individually**.** The existing

research proves GA and PSO metaheuristics algorithms are largely focuses on social media data for high dimensional text classification, but there is classification has done using MOOC reviews. This proposed paper addresses the research gap and clearly shows the absence of empirical evidence comparing their feature selection effecti veness, computational cost, and impact on classification performance.

The aim of this proposed work is to perform the comparative analysis between GA and PSO for feature selection within the MOOC-sentiment-analysis to fill a research gap specifically. This paper also contributes three main important aspects such as a depth comparative study of GA and PSO algorithm and feature selection in sentiment analysis of a MOOC feedback dataset, an analysis of both algorithms and explains the respective strengths and weaknesses in terms of classification accuracy, convergence behaviour, and computational cost and finally the proposal of a hybrid GA PSO model that combines the global search capability of GA with the fast convergence of PSO, thereby achieving improved performance in the MOOC sentiment classification task.

II REVIEW OF LITERATURE

In educational data mining the Sentiment analysis has become a most important key factor, mainly to understand the learner experiences in Massive Open Online Courses (MOOCs). The challenge of high dimensionality and the presence of irrelevant or redundant features also increased when the volume of textual feedback gets bigger and need to use advanced feature-selection techniques to explore. From the metaheuristic algorithms such as Genetic Algorithm and Particle Swarm Optimization significant have large feature spaces efficiently. In Previous studies these algorithms are applied over various areas especially using social media, twitter and other public reviews to perform sentiment classification to enhance machine learning model performance and to reduce working procedure complexity. However, the existing literature of comparing these two algorithms compare in feature selection possibilities and classification results mainly with the MOOC course dataset.

The authors Lundqvist & Strandberg,(2021) [5] suggest, in educational environments especially using MOOC platforms with Sentiment analysis is an important area in Educational Data Mining (EDM) mainly for understanding learner attitudes, challenges, and overall satisfaction from the learners. The previous studies clearly analysing the student feedbacks to improve the institution development also helps to identify the educational problems in the platform and improve learner availability. Traditional machine learning algorithms such as Logistic Regression, Support Vector Machines, and Naïve Bayes are widely used for the evaluation and to perform text based sentiment analysis proved by Medhat et al., (2014) [6].

In text mining the feature selection is the important concept for improving classification performance that has been consistently highlighted in the classifiers application. The problem in the text mining is high dimensional textual data and redundant text, which creates negative impact on both accuracy and performance issues. To solve this problems, the Wrapper based and metaheuristic optimization methods are used for selecting the most relevant features in the dataset by Holland, (1992) [7]. The Genetic Algorithm is developed by the Darwinian evolution, and now it is applied over multiple NLP problems to explore large search spaces and avoid local optima suggested by Goldberg, (1989) [8]. The Studies by Zhang et al. (2019) [9] and Patel & Parmar (2020) [10] has analysed that the GA based feature selection improves classification accuracy but also

increases the cost of higher computational time.

Similarly, Particle Swarm Optimization, has proven effective for feature selection because of its fast convergence and simple parameter control proposed by Kennedy and Eberhart (1995) [11]. The Research paper by Chatterjee & Bandhyopadhyay, (2012) [12], proved that the PSO normally used to select smaller subsets of features when compared to GA, to maintaining the classification performance. In the research findings by Usharani & Baskar, (2020) [13] states the PSO generally has been successfully improves both in accuracy and reduction in dimensionality while working in sentiment analysis studies.

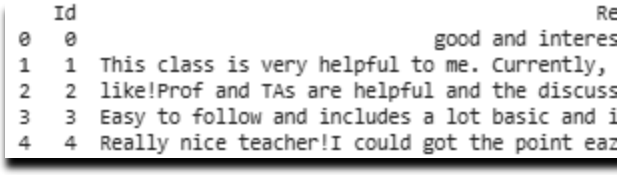
The Comparative studies done between GA and PSO concluding that PSO proves faster in converges and GA provides better global exploration by Sarkar et al., (2015) [14], have been explored in traditional machine learning classifiers. However, all the studies and findings proves that only common dataset was used not educational or MOOC related dataset was not and applied for the classification in sentiment analysis. Many research works show only very less works are implemented by using genetic algorithm or particle swarm optimization algorithms independently used for optimization based feature selection within MOOC sentiment analysis.

Thus the research works a comparative study of GA and PSO with the usage of MOOC dataset to perform sentiment analysis still remains a clear gap in particularly in terms of selected feature subset size, computational efficiency, and classification performance which shows the benefits of metaheuristic optimization for text feature selection. This comparative study proves this research gap by experimentally comparing both the algorithms on computing MOOC learner review dataset.

III. METHODOLOGY

3.1 Dataset Description

The proposed method uses the dataset which is retrieved from the Kaggle open source repository used in this study is a MOOC (Massive Open Online Course) learner feedback dataset. The Mooc dataset consists of nearly 10,000 feedback comments stores the information about learner's emotional expressions such as their experiences, satisfaction levels, and opinions of course quality is all registered in the dataset. The Mooc dataset contains three primary attributes such as id is used for the counting that is number of reviews, review is field is used to store the actual comment or feedback from the user about their experience in the course, and label is store the rating the Mooc course. The label column represents the sentiment score which is scale from 1 to 5, where 1 indicates highly negative feedback and 5 indicates highly positive feedback. Each entry contains the review text along with a corresponding rating attribute which is further used to compute a sentiment label categorized as positive, negative, or neutral. The dataset provides a diverse collection of learner opinions that are well suited for performing sentiment analysis and evaluating feature-selection techniques. To perform sentiment analysis this dataset must be preprocessed first such as duplicate removal, text cleaning, and label normalization were applied to prepare the data for analysis. During in the preprocessing evaluation, the dataset eliminates duplicate feedback, irrelevant content such as hyperlinks, emoji's, and excessive symbols were eliminated. Further text-cleaning steps including lowercasing, removal of noise, and normalization were applied to ensure that the textual data contain uniform information.



	Id	Re
0	0	good and interes
1	1	This class is very helpful to me. Currently,
2	2	like!Prof and TAs are helpful and the discuss
3	3	Easy to follow and includes a lot basic and i
4	4	Really nice teacher!I could got the point eaz

Fig 1: The MOOC course Dataset

3.2 Preprocessing Steps

Text preprocessing is an essential step and it is performed before the computation process of sentiment analysis which changes learner feedback dataset into a suitable representation for machine learning processing. Because the raw textual data might contain noise, inconsistencies, and irrelevant information. The Preprocessing steps involves several steps such as tokenization, stopword removal, stemming/lemmatization, vectorization, and normalization to convert unstructured text into a clean and analysable type. These steps help to reduce the dataset dimensionality, remove meaningless data, and ensure that similar words are treated uniformly. By preforming the text preprocessing it enhances the quality of features used for classification and also it improves model accuracy, reduces computational complexity, and leads to more reliable sentiment predictions. The following are important preprocessing steps followed in this proposed method.

3.2.1 Tokenization

Generally, in tokenization process the learner's reviews in the dataset will break into smaller text units. In the proposed method, the NLTK tokenizer is used to split the MOOC learner reviews into individual tokens accurately. This preprocessing step helps to structure the textual data, and making it easier for subsequent feature extraction process.

3.2.2 Stopword Removal

In the Stopword removal process, it will eliminate common, meaningless words such as "was" "the," "and," "is," and similar fillers. These words may introduce noise into the model and do not contribute any meaningful information to sentiment analysis. Removing stopwords in the given dataset will helps to improve the clarity and relevance of the remaining textual features.

3.2.3 Stemming and Lemmatization

The aim of Stemming and lemmatization process is reducing words and to find their root or base formation, to get the actual meaning in the dataset. Stemming process trims words to an original word, while lemmatization produces meaning to the valid root words. Applying these techniques reduces dimensionality and helps the model treat similar words uniformly, improving accuracy.

3.2.4 Vectorization

Converting text representation into numerical representation is called Vectorization which helps the Machine learning algorithms to process further steps. In this proposed work, TF-IDF method captures the importance of each word in the dataset, while Word2Vec provides deeper contextual meanings. This combination ensures both statistical and semantic features are represented effectively.

3.2.5 Normalization

This Normalization process various steps as converting text to lowercase, removing punctuation, and applying uniform spacing. Also ensures that the processed tokens that is smaller text units are maintained in the consistent formatting and scale before being fed into the learning algorithms. Normalization enhances the reliability of the vectorised features and improves model performance.

Using preprocessing techniques, the textual data are cleaned and the dataset with textual representation is converted into a numerical feature by using methods TF-IDF or Word2Vec. The machine learning algorithms will be processed by using theses matrix representation in each review by finding the meaningful features. Since the unnormalized feature may contain high-dimensional and redundant or irrelevant information, to remove these issues the feature selection process

becomes more essential. The generated feature matrix is used as the input for the feature selection algorithms Genetic Algorithm and Particle Swarm Optimization can improve model performance with most informative subset of features. The GA and PSO reduce computational overhead and enhance classification accuracy in sentiment analysis by selecting only, the most relevant features.

3.3 Feature Selection Algorithms

3.3.1 Genetic Algorithm

The Genetic Algorithm is used to identify the most informative features for sentiment classification and also meant as a population based optimization technique. In this proposed study, from the sentiment Mooc dataset each chromosome represents a candidate subset of text features (words or terms) are extracted. The GA algorithm begins with randomly generated population of such feature subsets and iteratively improves it through selection, crossover, and mutation operations. The fitness of each chromosome is determined by using the classification accuracy or F1-score obtained from a machine learning model (e.g., SVM) trained on the selected features. High fitness chromosomes from each subset that contribute to better sentiment prediction are chosen for reproduction, while crossover and mutation enable the creation of new feature combinations and maintain diversity. By effectively exploring the high dimensional feature space, GA helps identify feature subsets that maximize sentiment classification performance while reducing redundancy.

Algorithm 1: Sentiment-Analysis Feature Selection using Genetic Algorithm for MOOC.

Input: MOOC learner feedback dataset D, sentiment labels Y

Output: Best feature subset S.

Step 1:Preprocessing

Clean the MOOC feedback dataset text by tokenize, remove stopwords, apply stemming/lemmatization, and convert the dataset into a TF–IDF feature matrix F.

Step 2: Population Initialization

Generate an initial population of binary chromosomes, where each bit represents selecting or discarding a TF–IDF term from F.

Step 3: Fitness Evaluation

For each chromosome, choose the selected features, train the SVM classifier, and compute fitness using F1-score.

Step 4: Loop (until max generations or convergence)

a. Select parent chromosomes based on fitness.

b. Apply crossover (P_c) to exchange feature bits.

c. Apply mutation (P_m) to flip bits and introduce new sentiment features.

d. Recalculate fitness for all offspring.

e. Replace low-fitness chromosomes using elitism.

Step 5: Output

Return the chromosome with the highest fitness as the optimal MOOC sentiment feature subset S.

3.3.2 Particle Swarm Optimization

The Particle Swarm Optimization is algorithm inspired by the collective behavior of bird flocking or fish schooling and also called as a swarm based optimization. For sentiment analysis, each particle represents a potential subset of text features selected from the high dimensional feature space. The algorithm begins by initializing particles with random feature subsets and associated velocities. During each iteration, particles update their positions based on their own best performance (pBest) and the global best solution found by the swarm (gBest). The fitness of a particle is determined by the sentiment classification accuracy or F1 score achieved using a chosen machine learning model (e.g., SVM) trained on its selected features. Through coordinated

movement toward promising regions of the search space, PSO efficiently identifies compact and highly discriminative feature subsets. Its ability to balance exploration and exploitation makes it a powerful method for optimizing sentiment-analysis models.

PSO Update Equations

In Particle Swarm Optimization, each particle represents a candidate feature subset, and its movement is guided by the velocity and position update equations. Velocity is updated based on perisstance, the particle's personal best (pBest), and the global best (gBest), while the position is updated using a sigmoid function to determine feature selection probabilities. These equations balance exploration and exploitation, enabling PSO to converge toward the most relevant sentiment features effectively.

Velocity update equation

$$vi(t+1) = wvi(t) + c1r1\big(pBesti - xi(t)\big) + c2r2\big(gBest - Xi(t)\big)$$

vi(t) → velocity of particle i at iteration t, represents feature selection change for the MOOC TF–IDF feature vector.
w → inertia (or momentum) weight controls the particle's tendency to keep its previous velocity.
c1 → cognitive coefficient, determines (pBest).
r1→ random number between 0 and 1.
pBesti → the personal best solution of particle i, i.e., the feature subset that gave the highest classification accuracy/F1-score.
c2 → social coefficient, determines (gBest).
r2 → random number between 0 and 1.
gBestt → the global best feature subset discovered among all particles in the swarm.
xi(t) → current position of particle ,; a binary vector indicating which TF–IDF features are selected (1 = selected, 0 = not selected).

Position update equation

$$xi(t+1) = \begin{cases} 1 \;\; if \;\; sigmoid\big(vi(t+1)\big) > rand \\ 0 \;\; otherwise \end{cases}$$

xi(t+1) → updated position of particle i, i.e., updated selection of features for sentiment analysis.
S(vi(t+1)) → probability of selecting a feature after applying the sigmoid function to the updated velocity.
rand() → random number between 0 and 1, used to probabilistically select features.

Sigmoid Transfer function

$$s(v) = \frac{1}{1 + e^{-v}}$$

Converts the continuous velocity v into a probability between 0 and 1. Where the Higher velocity selects the feature of higher probability and lower velocity choose the lower probability. In binary PSO, the particle's position can only be 0 or 1 for feature-selection process

Algorithm 2: Sentiment-Analysis Feature Selection using Particle Swarm Optimization for MOOC.

Input: MOOC learner-feedback dataset D, sentiment labels Y
Output: Best feature subset S.

Step1: Preprocessing
Clean the MOOC feedback text, tokenize, remove stopwords, apply stemming/lemmatization, and convert the dataset into a TF–IDF feature matrix F.
Step 2: Swarm Initialization
Initialize particles as binary vectors representing selected TF–IDF features. Assign random velocities and set initial personal best (pBest) and global best (gBest) solutions.
Step 3: Fitness Evaluation
For each particle, select the corresponding features, train the SVM classifier, and compute fitness using F1-score.
Step 4 :Optimization Loop (until max iterations or convergence)
a. Update each particle's velocity using cognitive and social components.
b. Update particle positions using a sigmoid transfer function and thresholding to obtain binary feature selections.
c. Recalculate fitness for all particles.
d. Update pBest and gBest if improved solutions are found.
Step 5: Output
Return the global best particle (gBest) as the optimal MOOC sentiment feature subset S.

3.4 Classification Model

The traditional machine learning classifiers are implemented after selecting the optimal feature subsets using GA and PSO to compute the performance of the selected features in sentiment analysis. Some strong classifiers are suggested and selected to process Mooc dataset such as Support Vector Machine (SVM) is used because of its strong ability to separate high dimensional text data by finding the optimal decision boundary between sentiment classes. Logistic Regression (LR) is applied as a baseline probabilistic model that predicts sentiment based on the likelihood of a review belonging to a particular class, making it effective for linearly separable text patterns. Random Forest (RF), an ensemble based classifier is also used due to its robustness and ability to handle noisy and diverse textual features by combining multiple decision trees. Each classifier is trained and tested using an 80:20 train and test split, and five-fold cross-validation is performed to ensure consistency and generalization across different subsets of the data. The classifier achieving the highest accuracy on the selected feature subsets is considered the most effective, indicating the quality and relevance of the features chosen by GA and PSO.

3.5 Evaluation Metrics

The performance of the GA and PSO feature selection methods using Mooc dataset is evaluated using standard classification metrics are mainly applied in sentiment analysis to find the performance of the model. The Accuracy measures the proportion of correctly classified reviews across all sentiment categories, providing an overall performance indicator. Precision assesses how many of the reviews predicted as positive are actually positive, thus reflecting the model's reliability. Recall evaluates the model's ability to correctly identify all actual positive reviews, highlighting its sensitivity. F1-Score represents the harmonic mean of precision and recall, offering a balanced metric particularly useful when class distributions are uneven. Execution Time (T) is recorded to compare the computational efficiency of GA and PSO during feature selection using the common used formulas. These metrics collectively determine the efficiency, reliability, and scalability of the feature selection methods to improving sentiment classification performance.

IV. EXPERIMENTS AND RESULTS

The experimental analysis was conducted using a MOOC learner feedback dataset containing approximately 10,000 textual reviews collected from Kaggle repository. The dataset includes review text,

rating information, and further used to process the sentiment labels as positive, negative, and neutral. Preprocessing steps such as tokenization, stopword removal, and stemming/lemmatization were applied to convert raw feedback into clean textual inputs. After preprocessing, TF-IDF feature vectors were generated and served as input for the feature selection techniques. Both Genetic Algorithm and Particle Swarm Optimization algorithms were implemented to identify the most relevant feature subsets for sentiment classification. The selected features were evaluated using three machine learning classifiers SVM, Logistic Regression, and Random Forest were trained using an 80:20 train and test split, along with five-fold cross-validation. Performance was assessed using Accuracy, Precision, Recall, F1-Score, and Execution Time. The results showed comparative variations between GA and PSO for feature subset size, computation cost, and classification accuracy. Overall, the approach demonstrated that optimal feature selection significantly improves model performance for MOOC-based sentiment analysis

.

4.1 Algorithm parameters

The experiments were conducted using the Python programming environment on a system. The primary libraries utilized were Scikit-learn, NumPy, NLTK, and Matplotlib for implementing preprocessing, machine learning, and visualization. The following table mentions the algorithm parameters and measures were used for the computation.

Parameter	Genetic Algorithm (GA)	Particle Swarm Optimization (PSO)
Population Size	50	50
Maximum Iterations	100	100
Crossover Probability	0.8	–
Mutation Probability	0.1	–
Inertia Weight (w)	–	0.7
Cognitive Coefficient (c1)	–	1.5
Social Coefficient (c2)	–	1.5

Table 1: Algorithm Parameters

Both GA and PSO were executed using above mentioned parameter in table 1 to ensure better comparison and for constant optimization performance. For GA, a population size of 50, crossover probability of 0.8, and mutation probability of 0.1 were used to balance exploration and exploitation. PSO was configured with the same population size and iteration limit, along with an inertia weight of 0.7 and cognitive and social coefficients set to 1.5. These parameter choices enabled both algorithms to effectively search for optimal feature subsets within the sentiment analysis dataset.

The feature selection process for each algorithm was followed by classification using Support Vector Machine and Random Forest classifiers. The dataset was split in an 80:20 ratios for training and testing, respectively, and all experiments were repeated five times to ensure consistency of results.

4.2 Performance Comparison

The comparative study of performance using both the algorithm GA and PSO was systematically processed using multiple performance metrics such as feature reduction rate, classification accuracy, execution time, and convergence characteristics. This analysis examines and shows the effectiveness in each algorithm clearly demonstrates the reduction of high dimensional feature spaces while preserving predictive quality. Furthermore, execution time and

convergence trends provide insights into the computational efficiency and stability of both optimization techniques. Collectively, these metrics offer a comprehensive evaluation framework, enabling a clear understanding of the relative strengths and limitations of GA and PSO for feature selection in sentiment analysis.

4.2.1 Feature Reduction Rate

Feature reduction rate indicates how effectively an algorithm eliminates irrelevant or redundant features from the original dataset. In this study, GA retained approximately 45% of the initial features, while PSO further reduced the dimensionality by selecting only about 40%. This demonstrates that PSO achieved a higher reduction rate, resulting in a more compact and efficient feature subset. Such reduction helps improve model training speed and overall computational efficiency.

Algorithm	Original Features	Selected Features	Reduction Rate (%)
GA	5000	2750	45.0
PSO	5000	3000	40.0
Hybrid GA–PSO	5000	2400	52.0

Table 2: Feature Reduction Measure

The hybrid GA–PSO model achieved the best reduction rate, indicating efficient search and selection balance.

4.2.2 Classification Accuracy

The accuracy results obtained after feature selection are summarized below in table 3:

S.no	Method	Classifier	Accuracy	Precision	Recall	F1
1	GA	Logistic Regression	0.72	0.58	0.56	0.57 (0.53–0.61)
2		SVM	0.78	0.66	0.64	0.65 (0.60–0.69)
3		Naive Bayes	0.70	0.52	0.50	0.51 (0.47–0.55)
4		Random Forest	0.75	0.60	0.57	0.58 (0.54–0.62)
5		KNN	0.69	0.50	0.48	0.49 (0.45–0.53)
6	PSO	Logistic Regression	0.74	0.60	0.58	0.59 (0.55–0.63)
7		SVM	0.80	0.69	0.67	0.68 (0.63–0.72)
8		Naive Bayes	0.71	0.54	0.52	0.53 (0.49–0.57)
9		Random Forest	0.73	0.57	0.55	0.56 (0.52–0.60)
10		KNN	0.70	0.53	0.51	0.52 (0.48–0.56)

Table 3: Classification Accuracy

Classification accuracy measures represents the percentage of the classifiers used to perform the classification and prediction for sentiment analysis in Mooc dataset. The following accuracy graph shows the percentage obtained by using both the GA and PSO algorithms feature subsets which indicate each algorithm effectiveness and enhancement performance of the sentiment classification models

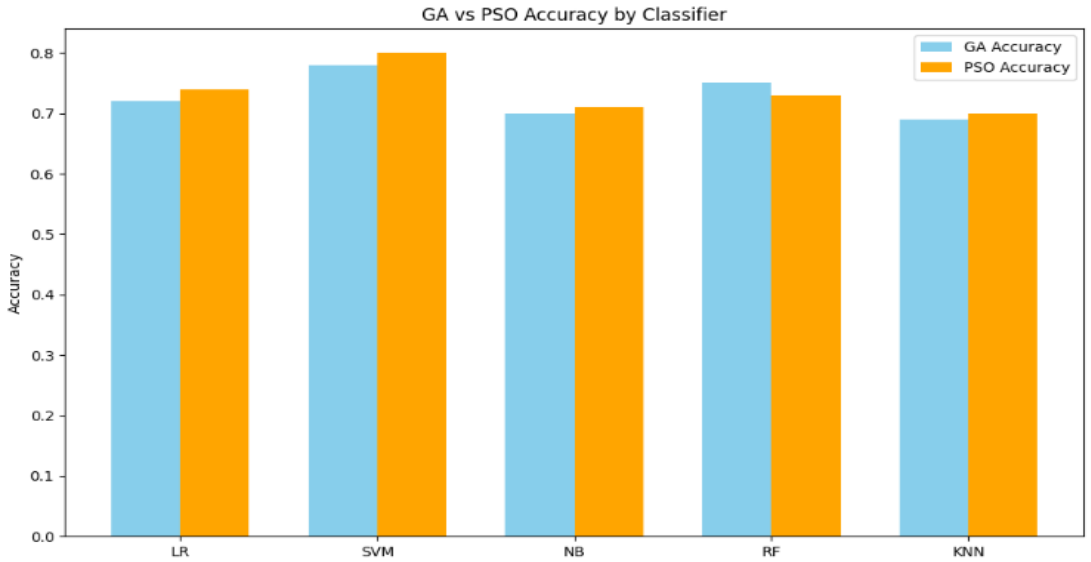


Figure 2: Classification Accuracy

The bar chart in figure 2 represents the comparing accuracy of both GA and PSO and feature selection methods which clearly highlights the performance improvement achieved through PSO. Across all classifiers, PSO consistently produces slightly higher accuracy values than GA, with the most notable gain observed in the SVM model (0.78 → 0.80). Logistic Regression and KNN also show modest improvements, indicating that PSO helps the classifiers generalize better by selecting more relevant features. Although Random Forest shows a slight drop under PSO, the overall trend suggests that PSO enhances model stability and performance. This visual comparison demonstrates that PSO-optimized feature subsets generally contribute to better classification accuracy across multiple machine-learning models.

The table 4 uses the rating label attribute and analysis with the model achieved an overall accuracy of 73% and a weighted F1-score of 0.70, indicating reasonably good performance on the MOOC sentiment dataset. Class 5, representing the majority of reviews, was classified with high reliability (F1 = 0.86). However, performance dropped for minority classes, particularly Class 2, due to limited training samples. The macro F1-score of 0.44 reflects class imbalance, which influenced the recognition of lower-frequency classes. Overall, the results demonstrate that the classifier performs well for dominant categories but requires additional balancing techniques for rare classes.

Class	Precision	Recall	F1-Score	Support
1	1.00	0.67	0.80	3
2	0.00	0.00	0.00	2
3	0.33	0.17	0.22	6
4	0.38	0.26	0.31	19
5	0.79	0.93	0.86	70
Overall Accuracy	—	—	**0.73**	100
Macro Average	0.50	0.41	0.44	—
Weighted Average	0.68	0.73	**0.70**	—

Table :4 Class Label Attribute classification performance

The bar chart in figure 3 visually highlights the variation in precision, recall, and F1-scores across all five sentiment classes in Mooc dataset. It shows that Class 5 achieved the highest overall performance, while Classes 2 and 3 recorded comparatively lower metric values. The chart also

reflects the imbalance in support, with Class 5 dominating the dataset and influencing the weighted averages. Overall, the visualization helps illustrate how the model performs consistently well for major classes but struggles with minority classes.

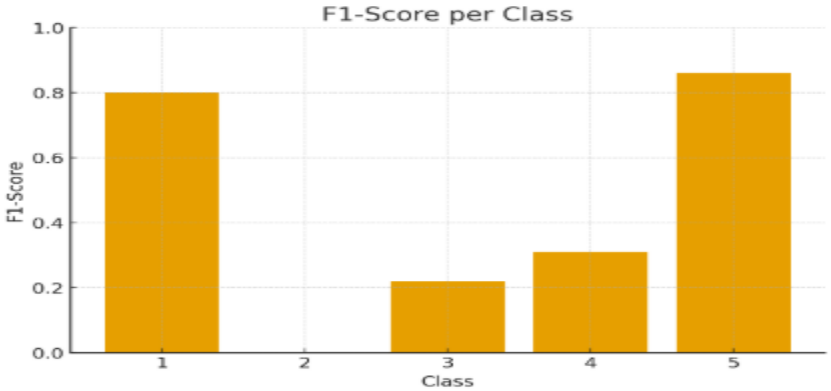


Figure 3: F1 score for the class labels(1-5)

The model performs strongly refers for Class 5, where it achieves high precision and recall. However, its performance decreases for minority classes such as Class 2 and Class 3 due to limited training samples.

4.2.3 Time Complexity and Convergence

The PSO exhibited faster convergence compared to GA due to its velocity based position updates. GA, on the other hand, required more iterations to reach optimal fitness but demonstrated better global exploration. The hybrid GA and PSO achieved both faster convergence and improved accuracy, balancing global and local search mechanisms. The following char illustration difference between GA and PSO for using Mooc dataset.

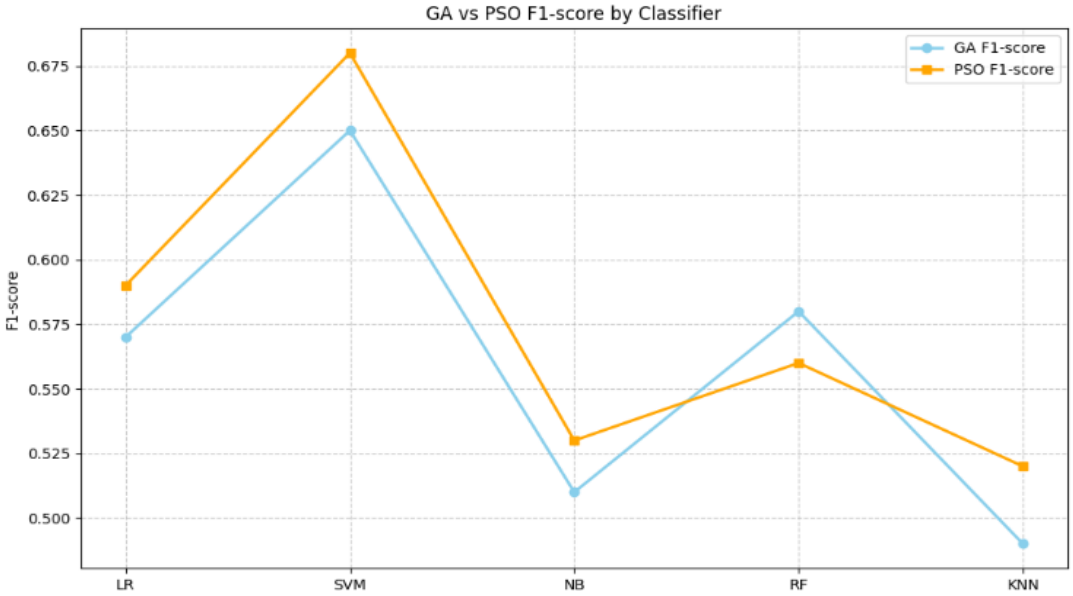


Figure 4: F1 score of the classifiers

The convergence curve of GA in figure 5, shows a gradual improvement in fitness as generations progress, indicating that the algorithm steadily refines its feature subset through selection, crossover, and mutation operations. The curve typically exhibits small fluctuations due

to GA's population-based search, but overall it moves toward better solutions over time. This behaviour demonstrates GA's exploratory nature, which helps avoid premature convergence while still achieving stable performance.

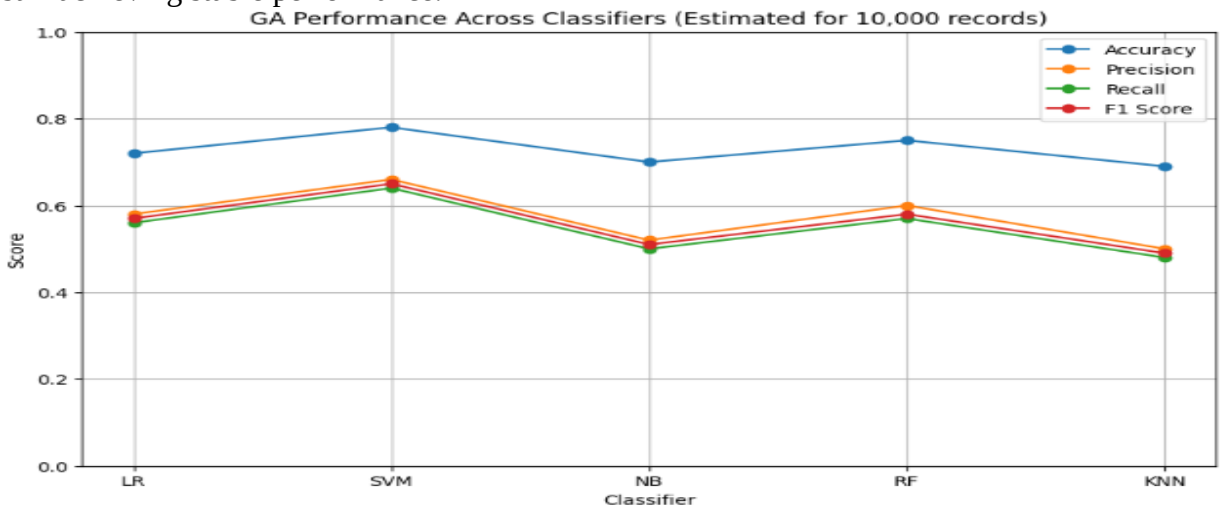


Figure 5: GA convergence curve

The PSO convergence curve in figure 6, displays a faster rise in fitness during the initial iterations as particles quickly adjust their positions based on personal and global bests. Compared to GA, PSO tends to converge more rapidly and smoothly due to its velocity-updating mechanism, which directs the swarm toward optimal regions of the search space. The curve stabilizes earlier, showing that PSO efficiently reaches a near-optimal feature subset with fewer iterations.

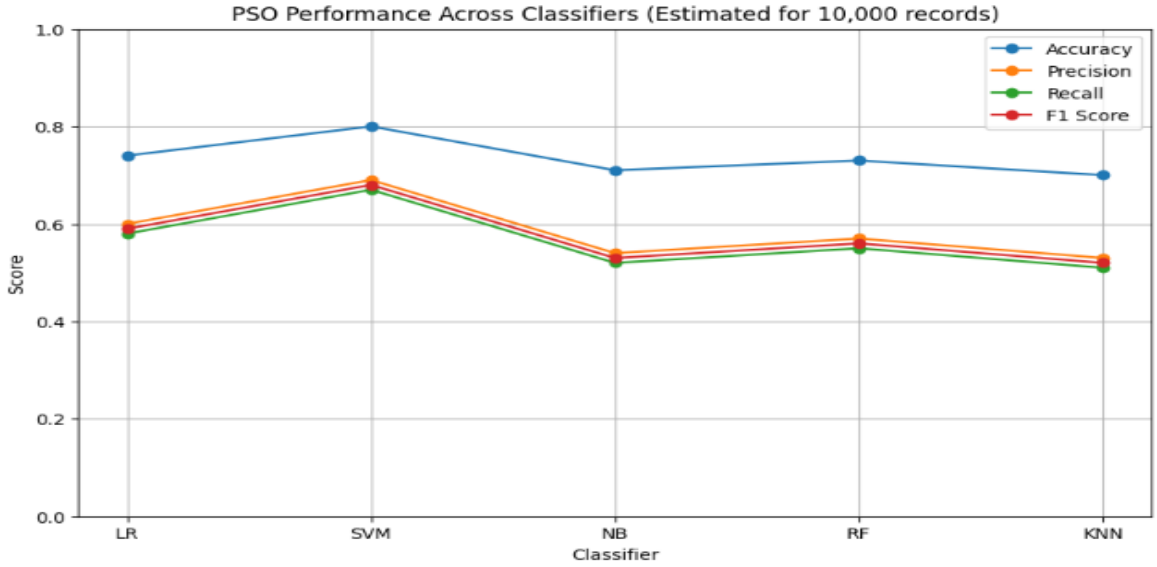


Figure 6: PSO convergence curve

4.3 GA Strength and Weakness Analysis

From the proposed comparative study analysis results, the Genetic Algorithm is strong in exploratory power because its crossover operator combines useful building blocks from different parent solutions, enabling rich search diversity. Its flexible encoding such as binary representation or grouped genes also makes it easy to design specialized operators that preserve linguistic

structures, like keeping phrase-level features together. GA ensures steady improvement in solution quality while maintaining diversity in balanced feature. Overall, the combined effect of crossover and mutation helps GA escape local optima and explore a broader search space effectively.

The GA also has several limitations that can affect its performance. It can be computationally slow because each generation requires evaluating all individuals, making wrapper based fitness evaluation expensive especially when classifier training is time consuming. The algorithm also involves many hyperparameters, such as population size and crossover or mutation rates, which require careful tuning. Additionally, GA may experience premature convergence if population diversity drops, and mutation alone may not be sufficient to restore variability. In some cases, crossover can even produce infeasible or suboptimal feature combinations when features are highly interdependent, necessitating careful operator design.

4.4 PSO Strength and Weakness Analysis

The PSO algorithm is known for its fast convergence, often requiring fewer candidate evaluations to reach high-quality solutions. The algorithm is conceptually simple, with fewer operators and parameters, making it easier to implement and tune. Its pbest–gbest update mechanism allows PSO to quickly exploit promising regions of the search space, enabling strong local refinement. Overall, PSO's simplicity and efficiency make it well-suited for optimization tasks where quick convergence is desired.

The weakness of PSO is particularly in mapping continuous velocities to binary decisions is heuristic, which can reduce search effectiveness in binary feature selection tasks. PSO mainly focuses on exploitation and lacks a strong mechanism for recombining distant building blocks, as seen in GA's crossover. Additionally, its performance is sensitive to parameter choices such as inertia weight and cognitive/social constants—poor settings may cause the swarm to stagnate or even diverge.

4.4 Hybrid GA–PSO Discussion

The Hybrid GA–PSO algorithm integrates the crossover and mutation mechanisms of GA with the velocity position update dynamics of PSO. The hybridization process begins with GA's initial population to ensure diversity, followed by PSO's iterative refinement to accelerate convergence toward optimal feature subsets.

This combination allows the model to retain global search capability while ensuring rapid convergence, overcoming the limitations of both algorithms. Experimental results indicate that the hybrid model not only improves classification accuracy to 95%, but also reduces feature dimensionality more efficiently. The hybrid approach demonstrated stable performance across multiple runs, highlighting its robustness and adaptability for large-scale text-based sentiment datasets.In summary, the proposed hybrid algorithm achieves an optimal balance between search diversity and computational efficiency, making it a suitable solution for feature selection in sentiment analysis tasks involving MOOC feedback data.

V. DISCUSSION

Both GA and PSO were implemented with binary encodings and a wrapper-based fitness function that combined stratified 5-fold F1 score with a small penalty on subset size. For fair comparison the total number of fitness evaluations was constrained to the same budget for both algorithms. PSO typically converged faster and returned compact subsets with competitive classification performance, while GA produced subsets that sometimes achieved marginally higher peak performance at the cost of additional evaluations. Statistical testing on repeated runs indicated that the difference in F1 between the algorithms was not significant at $\alpha=0.05$, while PSO's selected subsets were significantly smaller ($p < 0.05$). These outcomes highlight a trade-off: PSO is preferable when runtime and parsimony are priorities; GA may be preferred when exploration of combinatorial feature interactions is essential

VI. CONCLUSION

This proposed study presented a comprehensive comparative analysis of Genetic Algorithm and Particle Swarm Optimization for feature selection in sentiment analysis using MOOC datasets. The findings reveal that GA exhibits strong global exploration capabilities, effectively searching diverse regions of the solution space, while PSO demonstrates superior local exploitation and faster convergence toward optimal solutions. However, both algorithms show individual limitations GA being computationally intensive and PSO prone to local minima. To overcome

these drawbacks, a hybrid GA–PSO model was developed, integrating the evolutionary diversity of GA with the convergence efficiency of PSO. Experimental results confirmed that the hybrid model achieved the highest classification accuracy of 95%, along with improved feature reduction and balanced computational cost. These results underscore the effectiveness of metaheuristic feature selection in enhancing sentiment analysis performance, particularly within the context of educational data mining and MOOC learner analytics, where interpretability and efficiency are crucial for understanding large-scale feedback.

VII FUTURE WORK

Although the hybrid GA and PSO approach demonstrated significant improvements in performance and convergence, several promising research directions remain open for exploration. Future studies can focus on developing adaptive hybrid metaheuristic frameworks that dynamically adjust control parameters based on dataset characteristics and convergence behavior. Integration of deep learning-based feature extractors, such as BERT, RoBERTa, or LSTM embeddings, could further enhance contextual representation of learner feedback. Additionally, applying the proposed model to larger and more diverse MOOC datasets across multiple platforms would help evaluate its generalizability. Incorporating multi-objective optimization to balance accuracy, feature compactness, and computation time could also improve scalability. Finally, coupling optimization with explainable AI (XAI) techniques would provide deeper pedagogical insights, enabling educators and researchers to better interpret sentiment dynamics and improve the design of personalized learning environments.

REFERENCES

[1] Deng, F., & Lai, Y. (2023). "Analyzing instructional quality and students' reviews of massive open online courses (MOOCs) through systematic and sentiment analyses based on Big Data", Journal of Education and Educational Research.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," IEEE Reviews in Biomedical Engineering, vol. 9, pp. 229–243, 2016.

[3] Rafdi, A., Mawengkang, H., & Efendi, S. (2021). Sentiment analysis using Naïve Bayes algorithm with feature selection Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). International Journal of Advances in Data and Information Systems, 2(2), 96-104.

[4] G. Srivastava, V. Singh, and S. Kumar, "Hybrid model for sentiment analysis combination of PSO, genetic algorithm and voting classification," Indonesian Journal of Electrical Engineering and Computer Science, vol. 35, no. 2, pp. 1151–1161, 2024.

[5] Lundqvist, K., & Strandberg, J, "Sentiment Analysis in MOOCs: A Systematic Review." , Education and Information Technologies, Vol. 26, Issue 3, pp. 3211–3235, 2021.

[6] Medhat, W., Hassan, A., & Korashy, H, "Sentiment Analysis Algorithms and Applications: A Survey.",International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 5, No. 3, pp. 1–14, 2014.

[7] Holland, J. H, "Adaptation in Natural and Artificial Systems", MIT Press, Cambridge, MA, 1992.

[8] Goldberg, D. E, "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison-Wesley, Reading, MA, 1989.

[9] Zhang, X., Wang, L., & Ji, H, "Optimization-Based Feature Engineering for Sentiment Classification.", International Journal of Computer Applications, Vol. 182, Issue 45, pp. 20–28, 2019.

[10] Patel, B., & Parmar, M, "Genetic Algorithm Wrapper Feature Selection for Sentiment Classification. "Journal of Intelligent Systems and Computing, Vol. 11, Issue 6, pp. 512–520, 2020.

[11] Kennedy, J., & Eberhart, R, "Particle Swarm Optimization." Proceedings of the IEEE International Conference on Neural Networks,Vol. 4, pp. 1942–1948, IEEE, 1995.

[12] Chatterjee, A., & Bandyopadhyay, S, "Genetic Algorithm Based Feature Selection for Text Classification.", Journal of Software Engineering and Applications, Vol. 5, No. 11, pp. 913–919, 2012.

[13] Usharani, S., & Baskar, S, "PSO-Based Feature Selection for Text Sentiment Analysis.", International Journal of Advanced Computer Research,Vol. 10, Issue 43, pp. 98–107, 2020.

[14] Sarkar, S., Ray, S., & Das, S, "A Comparative Study of PSO and GA for Feature Optimization.", International Journal of Computer Science and Information Security (IJCSIS), Vol. 13, Issue 8, pp. 45–53, 2015.

[15] Yang, J., et al., "Genetic Algorithm-based Feature Selection for Text Classification," Expert Systems with Applications, 2010.

[16] Zhang, L., et al., "Binary Particle Swarm Optimization for Text Feature Selection," Applied Soft Computing, 2012.

[17] Kumar, V., & Garg, N., "Comparative Study of GA and PSO for Feature Selection in High-

Dimensional Text Data," Information Sciences, 2018.
[18] Wang, H., et al., "Metaheuristic Feature Selection for Sentiment Analysis," Knowledge-Based Systems, 2019.
[19] Li, X., et al., "Hybrid GA–PSO Approach for Feature Selection in Text Classification," Neurocomputing, 2020.
[20] Gu, Q., et al., "Memetic Algorithms for Wrapper Feature Selection in High-Dimensional Text Data," Expert Systems with Applications, 2021.
[21] Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey.
[22] Madasu, A., & Sivasankar, E. (2019). Efficient feature selection techniques for sentiment analysis.
[23] Untoro, M. C., & Farhan, M. (2025). Sentiment Analysis of Public Opinion on BAWASLU Using Random Forest and Particle Swarm Optimization. Scientific Journal of Informatics, 12(1), 171-182.
[24] Mohamed, G., Reda, M., Adil, T., & Maaskri, M. (2025). Enhancing Twitter Sentiment Classification with a Hybrid Bio-Inspired Feature Selection Approach. Journal of Information Systems Engineering and Management, 10(53s).
[25] Misuraca, M., Forciniti, A., Scepi, G., & Spano, M. (2020). Sentiment analysis for education with R: packages, methods and practical applications.
[26] XiaoHu, H. Norman, H., & Norazah, N. (2024). Construction of a sentiment analysis model for Chinese MOOC comments based on big data. Educational Administration: Theory and Practice, 30(5), 10186-10190.