Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing Twitter Sentiment Classification with a Hybrid Bio-Inspired Feature Selection Approach

Goismi Mohamed[1], Hamou Reda Mohamed[2] , Tomouh Adil[3] , Maaskri Moustafa[4]

[1] EEDIS Laboratory, Computer Science department, Djillali Liabes University, Sidi Bel Abbes, 22000, Algeria
E-mail: mohamed.goismi@univ-sba.dz
[2] GeCoDe Laboratory,Computer Science department, Dr.Tahar Moulay University, Saida, Algeria
E-mail: hamoureda@yahoo.fr
[3] EEDIS Laboratory, Computer Science department, Djillali Liabes University, Sidi Bel Abbes, 22000, Algeria
E-mail: toumouh@gmail.com
[4] Department of Electrical Engineering, Ibn Khaldoun -University, Tiaret, 14000, Algeria
E-mail: moustafa.maaskri@univ-tiaret.dz

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Sentiment analysis on Twitter is an essential task for extracting valuable information from public opinions. However, the highly dimensional and noisy nature of text data poses a challenge to achieving high classification accuracy. To address this problem, we propose in this study a hybrid feature selection approach that combines chi-square (a filtering method) with bio-inspired wrapper-based algorithms to improve classification performance. Specifically, we evaluate four hybrid approaches: Chi-Square + Genetic Algorithm (GA), Chi-Square + Particle Swarm Optimization (PSO), Chi-Square + Harris Hawks Optimization (HHO), and Chi-Square + Whale Optimization Algorithm (WOA), where wrapper methods refine features based on machine learning classifiers (KNM evaluator) followed by the use of classifiers (KNN, SVM, NB, RF, LR, MLP) as the base classifier. Experimental results show that the hybrid approach followed by the MLP classifier, achieve good results in terms of (Accuracy, Precision, Recall, and F1-score) for all used datasets (Sentiment140, IMDB, and Us airline tweets) outperforming simple approaches. The result is superior classification accuracy and better selection of feature subsets. These results highlight the effectiveness of integrating statistical filtering with bio-inspired optimization to improve sentiment analysis models by reducing computational complexity and improving predictive performance.

Keywords: Sentiment Analysis, Twitter, Filter Selection, Wrapper Selection, Feature Selection, Chi-Square, GA, PSO, HHO, WOA, Machine Learning.

INTRODUCTION

Social media sites like Twitter have developed into a vital source of up-to-date information, offering insightful data on the attitudes, trends, and opinions of the general population. The objective of sentiment analysis, often called opinion mining, is to categorize textual input into sentiments such as positive, negative, or neutral, which is a key issue in natural language processing (NLP) [1][2][35].

Accurate sentiment analysis of Twitter data has several uses, such as political analysis, brand monitoring, and consumer feedback evaluation [3][4]. Nevertheless, a number of issues with Twitter sentiment analysis, such as high dimensionality, noisy text, redundant features, and informal language, can have a detrimental effect on classification accuracy [5].

The process of feature selection is a key factor affecting the performance of sentiment classification models. Model overfitting and heightened computational complexity arise from the existence of redundant and non-essential features in high-dimensional feature spaces [6][7]. Conventional feature selection methods can be generally divided into filter techniques (which choose features according to statistical metrics like Chi-Square and Mutual Information) and wrapper techniques (which assess feature groups based on the performance of the model) [8]. Wrapper methods are more effective but require significant computational resources, whereas filter approaches are efficient in terms of computation yet overlook the relationships between features [9].

Hybrid feature selection strategies that blend filter and wrapper-based techniques have been investigated more and more in an effort to overcome these drawbacks. Specifically, bio-inspired algorithms have demonstrated great potential in feature selection optimization [10] [11]. These algorithms imitate natural selection, swarm intelligence, and predator-prey dynamics. These techniques reduce dimensionality and increase classification accuracy by effectively exploring huge feature spaces to find the most relevant subset [12][13].

In this work, we suggest a hybrid feature selection approach that combines the Chi-Square filter method with several bio-inspired wrapper-based optimization methods, such as Whale Optimization Algorithm (WOA), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Harris Hawks Optimization (HHO). To assess how well the suggested approach improves classification performance, it is used on the Sentiment140 dataset, a benchmark dataset for sentiment analysis on Twitter. Chi-Square + WOA and Chi-Square + HHO perform better than other hybrid approaches, attaining the maximum classification accuracy while preserving computing efficiency, according to our experimental data.

RELATED WORKS

Sentiment analysis on Twitter presents significant challenges due to the brevity, informality, and noisy nature of the content. Tweets often contain abbreviations, emojis, hashtags, and misspellings, leading to extremely high-dimensional feature spaces when employing bag-of-words, TF-IDF, or word embedding representations. Feature selection has thus become a critical step to reduce model complexity, improve learning performance, and avoid overfitting caused by noisy or redundant features. Several strategies have been employed for feature selection, including filter methods, wrapper methods, and metaheuristic optimization methods such as bio-inspired algorithms.

Filter methods score features individually based on statistical tests, independent of the classification algorithm. Among these, Chi-square (χ^2) feature selection has been widely adopted due to its computational efficiency and effectiveness. Chi-square tests evaluate the dependence between a feature and the class label, where higher χ^2 scores indicate stronger associations. Forman [8] conducted one of the first large-scale empirical studies on text classification and demonstrated that Chi-square selection achieved robust results [36], particularly for skewed datasets, outperforming simpler metrics such as document frequency. In the context of Twitter sentiment analysis, Saif et al. [14] showed that selecting the top 1,000 features based on Chi-square scores improved classifier accuracy by up to 7% compared to models trained on the full feature set. Similarly, Santos et al. [15] applied Chi-square feature selection for emoji-based sentiment classification on Twitter, finding that reducing the feature set by more than 80% resulted in less than a 2% decrease in model performance. Despite their speed and scalability, filter methods like Chi-square suffer from certain limitations, notably their inability to capture feature interactions and their tendency to select redundant features.

Wrapper methods, in contrast, evaluate feature subsets based on classifier performance, allowing them to consider feature interactions but at a high computational cost. Sequential Forward Selection (SFS) is one such approach, wherein features are added iteratively based on their impact on model accuracy. Siddiqua et al. [16] applied SFS alongside Support Vector Machines (SVMs) for Twitter sentiment analysis, achieving a 4–5% improvement in F1-scores compared to models without feature selection. Similarly, Silva et al. [17] compared wrapper-based and filter-based strategies, observing that wrappers outperformed filters by approximately 3% in F1-score on noisy Twitter datasets, albeit with a substantial increase in computational time. Although wrapper methods produce compact and effective feature subsets, their scalability remains problematic for large datasets.

To mitigate the weaknesses of both filters and wrappers, bio-inspired metaheuristic algorithms such as Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO) have been increasingly applied for feature selection. These algorithms offer a powerful balance between exploration and exploitation in vast search spaces and have proven capable of discovering globally optimal feature subsets. Camacho et al. [18][31] demonstrated that GAs could optimize feature subsets in a Twitter corpus, leading to an average accuracy improvement of 6% and a 70% reduction in feature dimensionality compared to non-optimized baselines. Similarly, Bermingham and Smeaton [19] applied GAs in combination with ensemble classifiers for sentiment polarity classification, emphasizing the advantages of maintaining diversity within optimized feature sets. PSO has also been widely explored for feature selection in sentiment analysis. Chakraborty et al. [18] proposed a PSO-based model initially designed for Amazon reviews but later adapted it to Twitter data, demonstrating better convergence rates and smaller selected feature sets relative to GA-based models. Further improvements were achieved by Aljarah et al. [21], who developed a hybrid filter-PSO model where Chi-square filtering was first applied to prune the feature space, thereby accelerating PSO convergence and boosting sentiment classification performance by 5–7%.

In addition to traditional bio-inspired algorithms like GA and PSO, more recent nature-inspired algorithms such as Harris Hawks Optimization (HHO) and Whale Optimization Algorithm (WOA) have been introduced for feature selection tasks. HHO, inspired by the cooperative hunting behavior of Harris hawks, offers strong global search capabilities and has been employed to optimize feature sets for sentiment classification tasks. Several studies, including that of Heidari et al. [22], demonstrated that HHO achieves competitive results compared to traditional metaheuristics, maintaining a good balance between exploration and exploitation. Similarly, WOA, which mimics the bubble-net hunting strategy of humpback whales, has shown promise in reducing feature dimensionality while preserving classification performance. Aljarah et al. [9] explored WOA for feature selection in Twitter sentiment analysis, demonstrating that WOA-based selection produced highly compact feature sets with competitive or superior classification accuracy compared to GA and PSO. Both HHO and WOA are particularly attractive due to their simplicity, fewer hyperparameters, and strong convergence behavior, making them suitable for complex high-dimensional datasets such as Twitter corpora.

Given the complementary strengths of filter and bio-inspired methods, hybrid approaches combining these techniques have gained increasing attention. Asghar et al. [23] proposed a two-phase feature selection strategy in which features were initially reduced using Chi-square selection and subsequently refined through Genetic Algorithms. Their hybrid approach yielded a 4% improvement in macro-averaged F1-score and reduced training times by approximately 30% compared to pure wrapper methods. Similarly, Habibi and Popescu [22] adopted a hybrid approach involving Information Gain filtering followed by Ant Colony Optimization (ACO) for fine-tuning, showing that the hybrid model consistently outperformed either method used independently, achieving at least a 5% performance gain. These results highlight the importance of balancing computational efficiency with the ability to uncover complex feature dependencies.

To evaluate feature selection methods for Twitter sentiment analysis, several standard datasets have been utilized. Sentiment140 [24], containing 1.6 million automatically labeled tweets, is the most commonly used large-scale benchmark. Other notable datasets include the Sanders Analytics Twitter Dataset, with 5,513 manually labeled tweets, and the SemEval competition datasets from 2013 to 2017, which cover a wide range of real-world sentiment classification challenges. Performance is typically assessed using metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC).

Despite notable progress, several challenges remain unresolved. Scalability is still a major concern, as wrapper and bio-inspired methods incur high computational costs on massive datasets such as Sentiment140. Additionally, noise in Twitter data — arising from typos, sarcasm, slang, and code-switching — complicates feature selection and classification. Most current approaches also primarily focus on textual features, whereas future work may need to integrate multi-modal features such as images, hashtags, and URLs for more robust sentiment analysis. Moreover, domain adaptation remains an open issue; feature sets optimized for one topic or domain (e.g., politics) often do not generalize effectively to others (e.g., entertainment or sports). Emerging strategies such as multi-objective optimization, which simultaneously considers classification accuracy and feature compactness, and deep feature selection techniques integrated into neural networks, represent promising avenues for future research.

BACKGROUND

Chi-Square (χ^2) for Feature Selection

Chi-square (χ^2) is a statistical method used in feature selection for classification tasks, particularly in text mining. It measures the dependence between a feature and the class label by calculating the divergence between observed and expected frequencies of feature occurrences. Features with high χ^2 scores indicate strong associations with the class and are selected for model training. It is widely used in filter methods due to its computational efficiency and simplicity[36]

Mathematically, the chi-square statistic for a feature *f* and a class label *c* is given by:

$$x^2(f,c) = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{1}$$

Where $O_{i,j}$ is the observed frequency and $E_{i,j}$ is the expected frequency for feature *f* and class *c*.

Particle Swarm Optimization (PSO)

PSO is a nature-inspired optimization algorithm based on the social behavior of birds flocking or fish schooling. It was proposed by Kennedy and Eberhart in 1995 [29] and is widely used for continuous and discrete optimization problems, including feature selection in machine learning. PSO optimizes a population of candidate solutions, known as particles, which explore the search space by updating their positions and velocities based on their personal best position and the global best found by the swarm.

In feature selection, every particle symbolizes a potential solution, with each dimension relating to a feature within the dataset. The algorithm looks for the optimal set of features that enhances the performance of a classification model.

Mathematical formulation :

$$v_i(t+1) = wv_i(t) + c_1r_1(p_{best,i} - x_i(t)) + c_2r_2(g_{best} - x_i(t)) \tag{2}$$

Where:

- $v_i\left(t+1\right)$ is the velocity of particle *i* at time *t+1*,
- $x_i\left(t\right)$ is the position of particle *i* at time *t*,
- $p_{best,i}$ and g_{best} are the personal best and global best positions,
- *w* is the inertia weight,
- c_1,c_2 are cognitive and social coefficients,
- r_1,r_2 are random numbers.

Genetic Algorithm (GA)

Genetic Algorithms (GAs) are a class of optimization algorithms inspired by the process of natural selection and genetics. Introduced by Holland in 1975 [30], GAs mimic the process of evolution, where a population of potential solutions evolves over generations through selection, crossover, and mutation operators. In feature selection, a chromosome represents a subset of features, and the fitness of the chromosome is evaluated based on the classification accuracy or other relevant metrics.

The process of evolution includes choosing individuals based on their fitness, merging them via crossover, and sometimes adding mutations to investigate new regions of the search space. Gas are particularly useful for solving combinatorial optimization problems, like feature selection, where the goal is to find an optimal subset of features.

Mathematical formulation:

- **Crossover**: Combining two parent solutions to create offspring.
- **Mutation**: Randomly altering an offspring to maintain diversity.
- **Fitness**: Evaluating the quality of a solution (e.g., classification accuracy)

Harris Hawks Optimization (HHO)

Harris Hawks Optimization (HHO) is a recently introduced metaheuristic optimization algorithm inspired by the hunting behavior of Harris's hawks. HHO utilizes various strategies for exploring and exploiting the search space, including surprise pounce (ambush), soft and hard besiege, and the player–prey dynamic. This makes HHO a versatile optimization tool that balances global exploration and local exploitation, making it effective in high-dimensional and complex optimization problems, such as feature selection.

The main advantage of HHO over other algorithms is its dynamic and adaptive search behavior, which increases the probability of finding the optimal or near-optimal solution.

Whale Optimization Algorithm (WOA)

The Whale Optimization Algorithm (WOA) was proposed by Mirjalili and Lewis in 2016 and is inspired by the bubble-net hunting technique of humpback whales. During hunting, humpback whales create a bubble-net to trap fish, and WOA mimics this behavior for searching and exploiting the optimal solution in optimization problems. WOA is characterized by its simple structure, fewer hyperparameters, and strong global search capability.

In feature selection, WOA searches the feature space to identify the most informative subset of features that improves classification performance. The algorithm uses the concepts of exploration (searching around) and exploitation (refining solutions) effectively to select the best features.

Mathematical formulation:

$$X(t+1) = X_{best} - A \cdot D \tag{3}$$

Where:

- X_{best} is the current best solution,
- A and D are coefficients that control the search dynamics.

METHODS

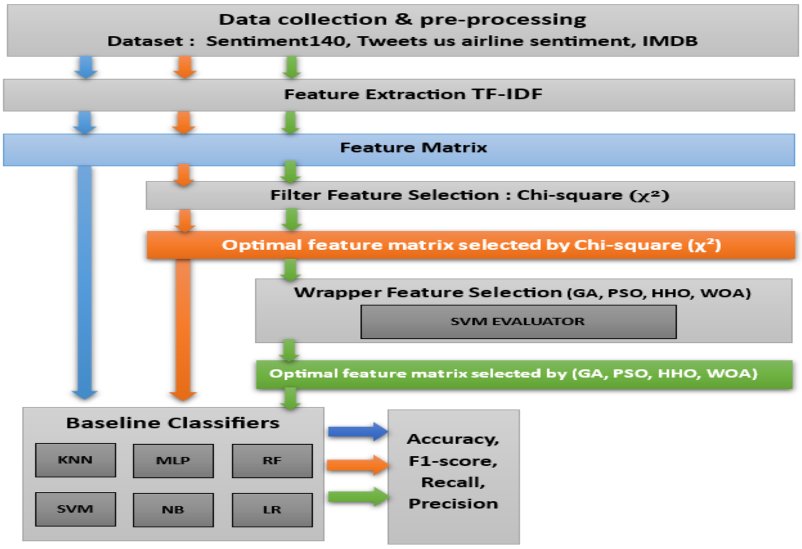
Data collection & pre-processing
Dataset : Sentiment140, Tweets us airline sentiment, IMDB
Feature Extraction TF-IDF
Feature Matrix
Filter Feature Selection : Chi-square (χ²)
Optimal feature matrix selected by Chi-square (χ²)
Wrapper Feature Selection (GA, PSO, HHO, WOA)
SVM EVALUATOR
Optimal feature matrix selected by (GA, PSO, HHO, WOA)
Baseline Classifiers
KNN
MLP
RF
SVM
NB
LR
Accuracy,
F1-score,
Recall,
Precision

Figure 1. Proposed systematic workflow for optimized sentiment classification

The figure **Figure 1** presents a comprehensive and systematic machine learning pipeline designed for sentiment analysis, leveraging three datasets including **Sentiment140**, **Tweets US Airline Sentiment,** and **IMDB** reviews. The process begins with data collection and pre-processing, where textual data is cleaned and normalized to remove noise, such as special characters, URLs, and stopwords, ensuring high-quality input for subsequent stages.

Feature extraction is performed using TF-IDF (Term Frequency-Inverse Document Frequency), which transforms the pre-processed text into a numerical feature matrix. This matrix captures the significance of words relative to their frequency in the corpus, providing a structured representation for machine learning models. To enhance model efficiency and reduce dimensionality, feature selection is applied using two distinct methods. The filter-based approach employs the Chi-square (χ2) test to identify the most statistically relevant features, generating an optimal feature matrix. This matrix is then used to train a suite of baseline classifiers, including KNN (with K=3 and K=4), SVM, Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Multilayer Perceptron (MLP). Additionally.

For further optimization, wrapper-based feature selection methods are applied, utilizing metaheuristic algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Harris Hawks Optimization (HHO), and Whale Optimization Algorithm (WOA). These algorithms iteratively search for the best feature subsets, with SVM serving as the evaluator to guide the selection process. The resulting optimal feature matrices from these methods are then evaluated alongside the Chi-square-selected matrix using a comprehensive set of metrics: accuracy, F1-score, recall, and precision.

This pipeline highlights a rigorous approach to sentiment analysis, combining traditional feature selection techniques with advanced optimization algorithms to maximize model performance. The inclusion of diverse datasets and multiple evaluation metrics ensures robustness, making the framework adaptable to various text classification tasks. By systematically comparing filter and wrapper methods, the pipeline provides valuable insights into feature engineering strategies, ultimately aiming to enhance the accuracy and reliability of sentiment analysis models in real-world applications.

1. DATASETS

This research employs three sentiment analysis datasets: Twitter US Airline Sentiment, Sentiment140, and IMDb. The IMDb dataset, initially gathered from the IMDb platform, was made publicly available by Maas et al. [25]. It includes two columns: review and sentiment. The dataset comprises 50,000 film reviews, equally divided into 25,000 positive and 25,000 negative entries. You can access it at:
https://www.kaggle.com/lakshmi25npathi/imdbdataset-of-50k-movie-reviews
The US Airline Sentiment dataset on Twitter contains 14,640 messages categorized into three labels: positive, negative, and neutral. The dataset is unbalanced, featuring 9178 negative reviews, 2363 positive reviews, and 3099 neutral reviews. The dataset includes tweets associated with six U.S. airlines: United, US Airways, Southwest, Delta, and Virgin America. The aim of collecting this data was to analyze the feelings of customers from every airline. The dataset comprises ten columns, including tweet ID, sentiment, text, airline name, and additional details. Nonetheless, only the text and sentiment columns will be utilized for sentiment analysis
The Sentiment140 dataset is evenly distributed, containing 0.8 million positive and 0.8 million negative sentiments. The dataset comprises six columns: target, id, text, flag, user, and id. In the experiment, the columns for the target and text are utilized for training and evaluation. Stanford University [26] acquired the dataset from Twitter. It can be accessed at https://www.kaggle.com/kazanova/sentiment140
The IMDb and Sentiment140 datasets have almost an equal number of samples for both positive and negative categories. Nonetheless, the Twitter US Airline dataset is uneven, with the negative class significantly outnumbering the positive and neutral classes. Considering this, a downsampling approach is utilized for the Twitter US Airline Sentiment dataset. Following downsampling, the sample sizes in each of the three classes are equal and evenly distributed. Figure 3 contrasts the sample distributions of the IMDb dataset both with and without data augmentation, along with the Twitter US Airline Sentiment dataset prior to and following downsampling.

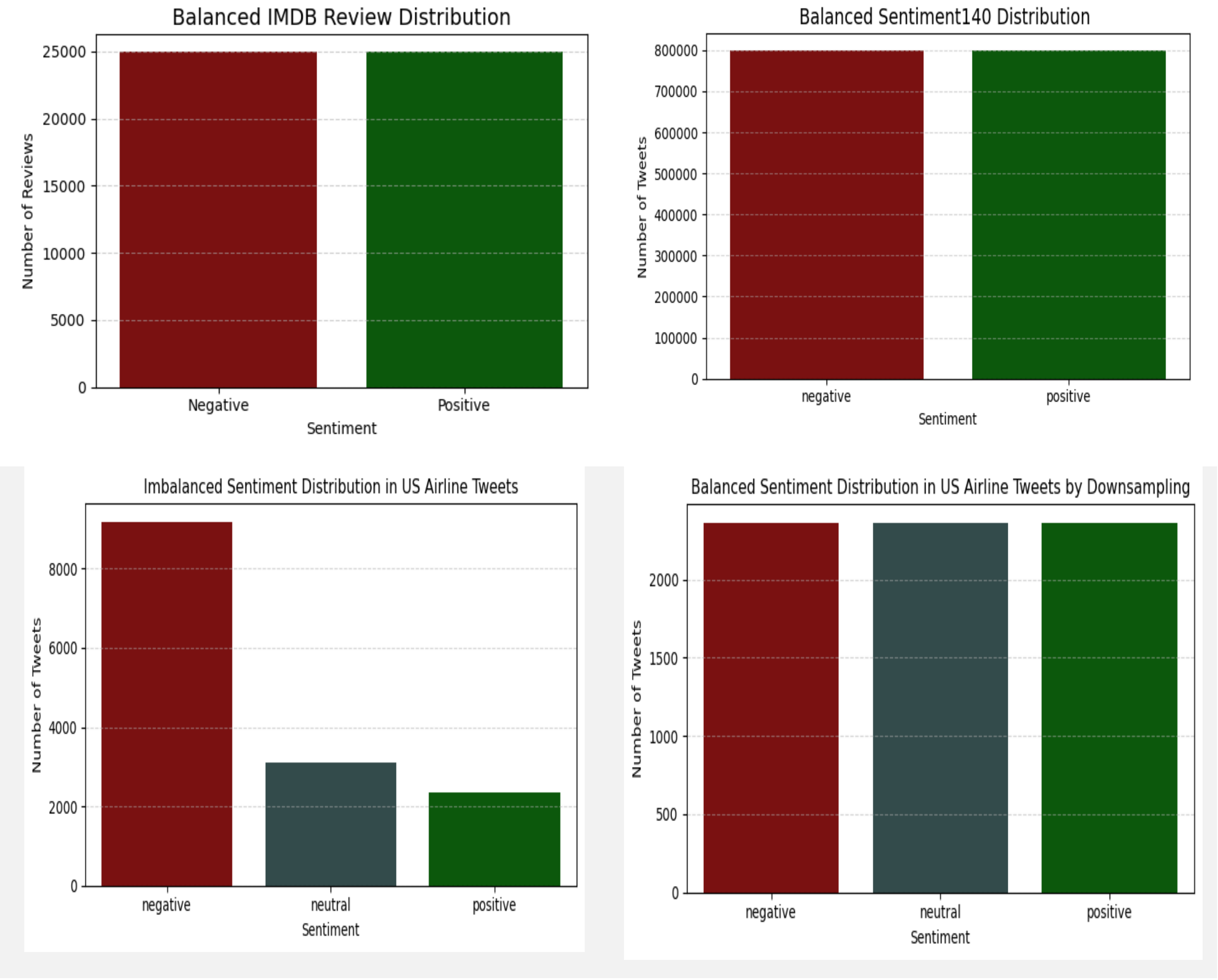

Figure 2. Class distribution of datasets

2. PREPROCESSING

To prepare the data for the extraction of pertinent features, pre-processing is carried out [27]. It entails removing noise from the dataset. Here, "noise" essentially refers to the grammatical errors that are frequently seen in any microblog post. The data is transformed to structured input format since the unstructured and noisy data degrades the sentiment classification task's quality [28]. The following tasks are part of the pre-processing approach used in this study:

- Elimination of all URL links, non-ASCII and non-English characters, digits, and stop words such as "the," "at," "as," "of," "to," and so forth.
- Negative mentions are transformed into their original, meaningful forms using the Internet Slang Dictionary [20], such as replacing contractions like 'couldn't' with 'could not', and expanding acronyms and slang expressions.
- Emoticons are replaced with their corresponding textual representations using an emoticon dictionary.
- Stemming and tokenization were applied to the text. Stemming reduces words to their root forms—for example, 'helping' becomes 'help'—to ensure consistency. Tokenization involves breaking the text into individual words or sequences of words (n-grams). This step was performed using the Tweet-NLP tool [34].

Take a look at this sample tweet from the Sentiment140 dataset:
" @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D "

Following the application of pre-processing, the following result was produced:

" awww bummer get david carr third day "

3. FEATURE EXTRACTION

This step finds the dataset properties that are most useful for detecting sentiments. The primary goal is to identify distinct features that can divide the data into positive, negative, and neutral classifications, resulting in increased sentiment classification accuracy. The feature matrix is generated using the term-frequency-inverse document frequency (TFIDF) algorithm [32][33]. This numerical statistic reflects a word's importance inside the specified corpus. **Table 1** displays the TFIDF score matrix for the given pre-processed sample tweet.

Table 1: *TF-IDF score matrix for the given pre-processed sample tweet*

awww	bummer	carr	david	day	get	third
0.577	0.408	0.288	0.288	0.288	0.577	0.288

4. FILTER FEATURE SELECTION: CHI-SQUARE (χ^2)

Formula (1) shows the Chi-Square value assigned to each feature or word in the dataset, which is the outcome of feature selection using the Chi-Square approach. The top 50% of features from the original feature matrix were chosen because they were associated with the target class. This method made it easier to find the most instructive characteristics for categorization and data analysis[36]**. Table 2** shows the Chi-square feature selection score matrix for the given pre-processed sample tweet.

Table 2: Chi-square feature selection score matrix for the given pre-processed sample tweet

awww	bummer	carr	david	day	get	third
0.254	0.183	0.001	0.009	0.450	0.021	0.183

5. WRAPPER FEATURE SELECTION

Wrapper methods help identify the most important features by combining search or optimization algorithms with machine learning models. Unlike other approaches, wrapper methods work closely with a classifier and use a fitness function—usually based on classification accuracy—to evaluate different feature subsets. Essentially, they integrate feature selection directly into the learning process to find the most relevant features.

These methods are widely used because they often lead to better classification performance and help reduce the number of unnecessary features. Here's how it works: first, a machine learning model evaluates different feature combinations using a fitness score, which reflects how well the model performs with those features. The process continues iteratively, searching for the feature set that delivers the best results. Once the optimal features are identified, they are used for the final text classification task.

In this step, we used the PSO, GA, HHO, and WOA methods with SVM as the evaluation algorithm, using the parameters shown in **Table 3**.

Table 3: *PSO, GA, HHO, and WOA parameters*

<table>
<tr><td colspan="5">PSO Parameters</td><td colspan="4">GA Parameters</td></tr>
<tr><td>Number of particles</td><td>Number of iterations</td><td>Inertia weight w</td><td>Acceleration factor C1</td><td>Acceleration factor C2</td><td>Number of chromosomes</td><td>Number of iterations</td><td>Mutation rate</td><td>Crossover rate</td></tr>
<tr><td>60</td><td>50</td><td>0.7298</td><td>1.496</td><td>1.496</td><td>60</td><td>50</td><td>0.2</td><td>0.9</td></tr>
<tr><td colspan="5">HHO Parameters</td><td colspan="4">WOA Parameters</td></tr>
<tr><td>Population</td><td colspan="2">Number of iterations</td><td colspan="2">Constant beta</td><td>Population</td><td colspan="2">Number of iterations</td><td>Constant beta</td></tr>
<tr><td>60</td><td colspan="2">50</td><td colspan="2">1.5</td><td>60</td><td colspan="2">50</td><td>1</td></tr>
</table>

6. BASELINE CLASSIFIERS

Six basic algorithms were used for the final classification: K-nearest neighbor (K=3 & k=4), Multi-layer Perceptron, Random Forest, Support Victor Machine, Naives Base, and Logistic Rregression.

RESULTS AND DISCUSSION

Table 4: Results of Sentiment140 Dataset

Without FS Sentiment140 dataset					With Chi2 only Sentiment140 dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	57,39	79,61	13,42	22,97	KNN K=3	74,21	85,35	54,96	66,86
KNN K=4	57,04	87,36	10,80	19,23	KNN K=4	70,76	88,14	44,17	58,85
MLP	83,77	84,00	81,18	82,57	MLP	87,08	86,01	86,83	86,42
NB	82,83	83,86	78,91	81,31	NB	86,67	87,29	84,07	85,65
SVM	85,51	87,13	81,41	84,17	SVM	87,65	89,02	84,29	86,59
RF	77,85	87,67	61,90	72,57	RF	82,74	85,98	75,91	80,63
LR	86,21	88,20	81,80	84,88	LR	87,86	89,18	84,62	86,84
With Chi2+PSO Sentiment140 dataset					With Chi2+GA Sentiment140 dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	85,71	90,52	78,01	83,80	KNN K=3	84,91	90,13	76,51	82,76
KNN K=4	81,42	92,81	65,84	77,04	KNN K=4	80,88	92,37	64,97	76,29
MLP	93,20	94,03	91,44	92,72	MLP	91,34	92,61	88,79	90,66
NB	86,68	87,90	83,32	85,55	NB	85,3	87,24	80,76	83,87
SVM	86,59	88,73	82,10	85,29	SVM	85,68	88,58	80,07	84,11
RF	85,93	90,67	78,34	84,06	RF	85,03	91,10	75,79	82,74
LR	85,41	88,24	79,80	83,81	LR	84,41	87,89	77,77	82,52
With Chi2+WOA Sentiment140 dataset					With Chi2+HHO Sentiment140 dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	83,45	92,07	71,15	80,27	KNN K=3	84,26	91,97	73,13	81,47
KNN K=4	77,48	93,77	56,15	70,24	KNN K=4	77,65	93,70	56,58	70,55
MLP	97,42	97,52	97,01	97,27	MLP	97,45	97,23	97,38	97,30
NB	88,25	88,97	85,82	87,37	NB	88,55	88,80	86,75	87,77
SVM	89,65	90,70	87,06	88,84	SVM	89,64	90,65	87,10	88,84
RF	87,23	92,90	79,06	85,42	RF	86,65	92,57	78,07	84,70
LR	87,64	89,90	83,24	86,4	LR	87,53	89,78	83,11	86,32

The results -Table 4 - clearly demonstrate that feature selection has a strong positive impact on model performance. Without any reduction, certain models—like KNN—struggle, especially when it comes to recall and F1-score. Even just applying the Chi-square (Chi2) method brings noticeable improvements, particularly for models like KNN, Naive Bayes (NB), and Random Forest (RF).

That said, the best results come from combining Chi2 with optimization algorithms such as PSO, GA, WOA, and HHO. The MLP model, for example, achieves outstanding results—reaching over 97% accuracy when paired with WOA and HHO. Other models like SVM, RF, and Logistic Regression (LR) also show significant improvements across precision, recall, and F1-score. PSO consistently performs well across all models, while WOA and HHO prove especially effective for more complex models.

In short, integrating Chi2 with optimization algorithms doesn't just enhance classification accuracy—it also helps models become more reliable and robust, particularly when working with complex datasets like Sentiment140.

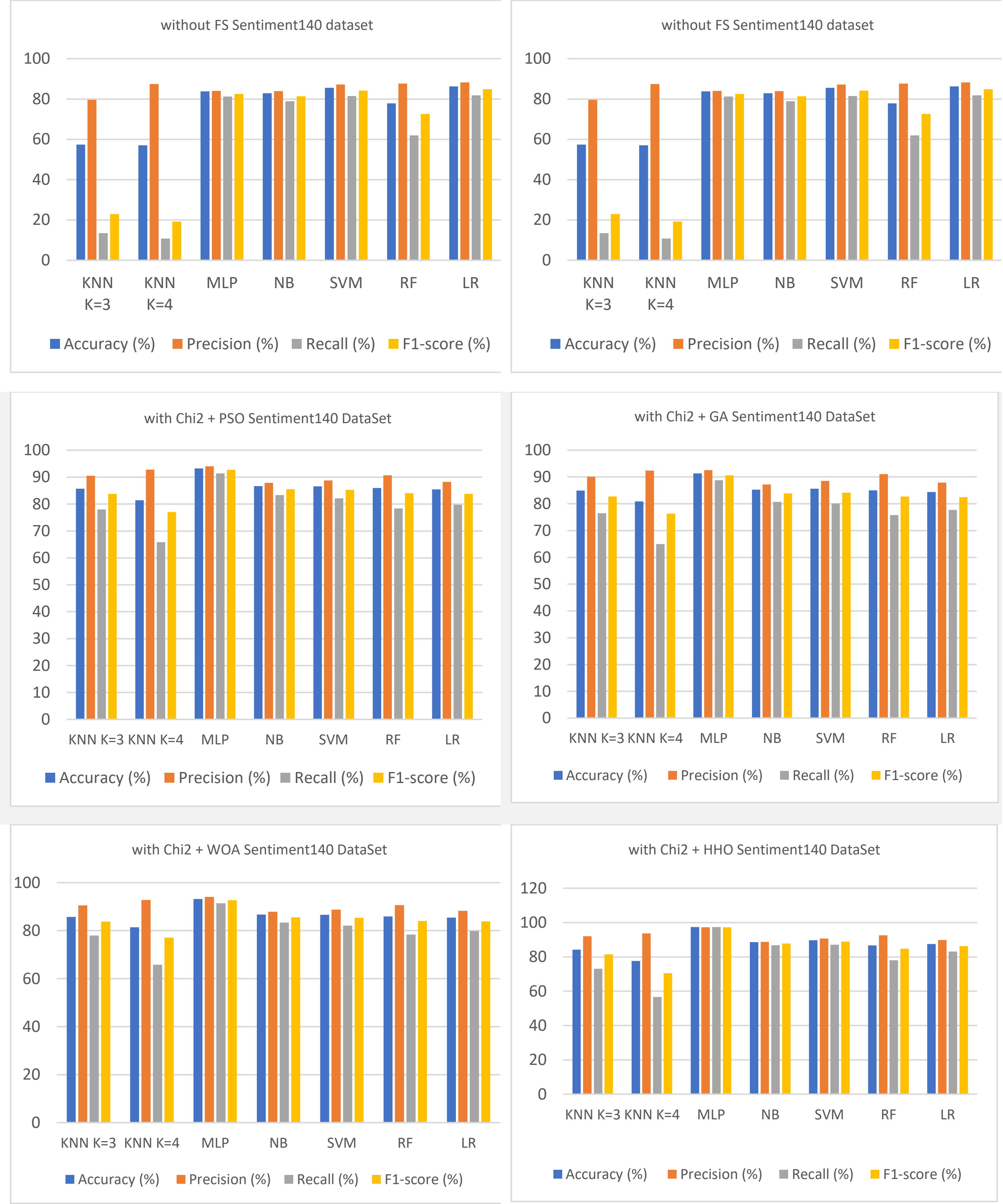

Figure 3: Accuracy, Precision, Recall, and F1-score of proposed model for all algorithms using Sentiment140 dataset

Table 5: Results of IMDB Dataset

Without FS IMDB dataset					With Chi2 only IMDB dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	70,7	75,15	83,65	79,17	KNN K=3	66,66	66,68	89,81	79,95
KNN K=4	69,23	78,28	74,43	79,31	KNN K=4	66,99	67,02	89,27	80,02
MLP	82,78	86,23	88,22	87,21	MLP	83,69	86,27	89,80	88,01
NB	75,84	78,99	86,80	82,71	NB	80,03	85,46	84,35	84,90
SVM	84,67	87,06	90,40	88,70	SVM	85,85	87,68	91,62	89,61
RF	71,67	70,86	97,58	82,10	RF	76,19	75,83	94,29	84,06
LR	85,22	86,97	91 ,51	89,18	LR	85,96	87,60	91,91	89,70
With Chi2+PSO IMDB dataset					With Chi2+GA IMDB dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	85,39	87,74	90,73	89,21	KNN K=3	59,82	91,54	43,68	59,15
KNN K=4	82,91	91,75	81,67	86,42	KNN K=4	49,32	92,21	26,08	40,67
MLP	98,09	98,52	98,61	98,57	MLP	98,06	98,53	98,56	98,54
NB	80,97	86,11	85,15	85,63	NB	80,84	85,67	85,53	85,60
SVM	86,57	88,15	92,23	90,14	SVM	86,46	88,06	92,16	90,03
RF	91,57	89,39	90,11	91,00	RF	91,54	89,21	89,15	93,92
LR	83,74	84,47	92,60	88,35	LR	83,86	84,54	92,72	88,44
With Chi2+WOA IMDB dataset					With Chi2+HHO IMDB dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	68,97	68,24	89,89	81,08	KNN K=3	68,17	67,67	89,90	80,69
KNN K=4	69,29	68,49	89,79	81,23	KNN K=4	70,82	69,57	89,86	82,00
MLP	98,49	98,60	99,12	98,87	MLP	98,65	98,65	98,86	98,11
NB	81,37	87,00	84,66	85,82	NB	81,20	87,00	84,36	85,81
SVM	90,40	91,67	94,12	92,88	SVM	90,24	91,49	94,09	92,77
RF	91,75	89,39	89,41	94,13	RF	92 ,01	89,74	89,35	94,31
LR	86,99	87,78	93,48	90,53	LR	86,62	87,55	93,69	90,51

Experiments on the IMDB dataset - **Table 5-** demonstrate that feature selection enhances classification algorithms' overall performance. Chi2 by itself yields modest increases, but the greatest outcomes are obtained when combined with optimization methods like PSO, GA, WOA, or HHO. PSO is notable for its excellent performance among these combinations, especially for models like KNN and MLP. While GA works well with models like MLP, it is less appropriate for KNN. In contrast, WOA and HHO also produce great results, closely matching each other. With many approaches, MLP achieves about 99% accuracy, making it the best-performing model overall. These findings support the usefulness of integrating optimization strategies with statistical selection techniques to enhance classification quality.

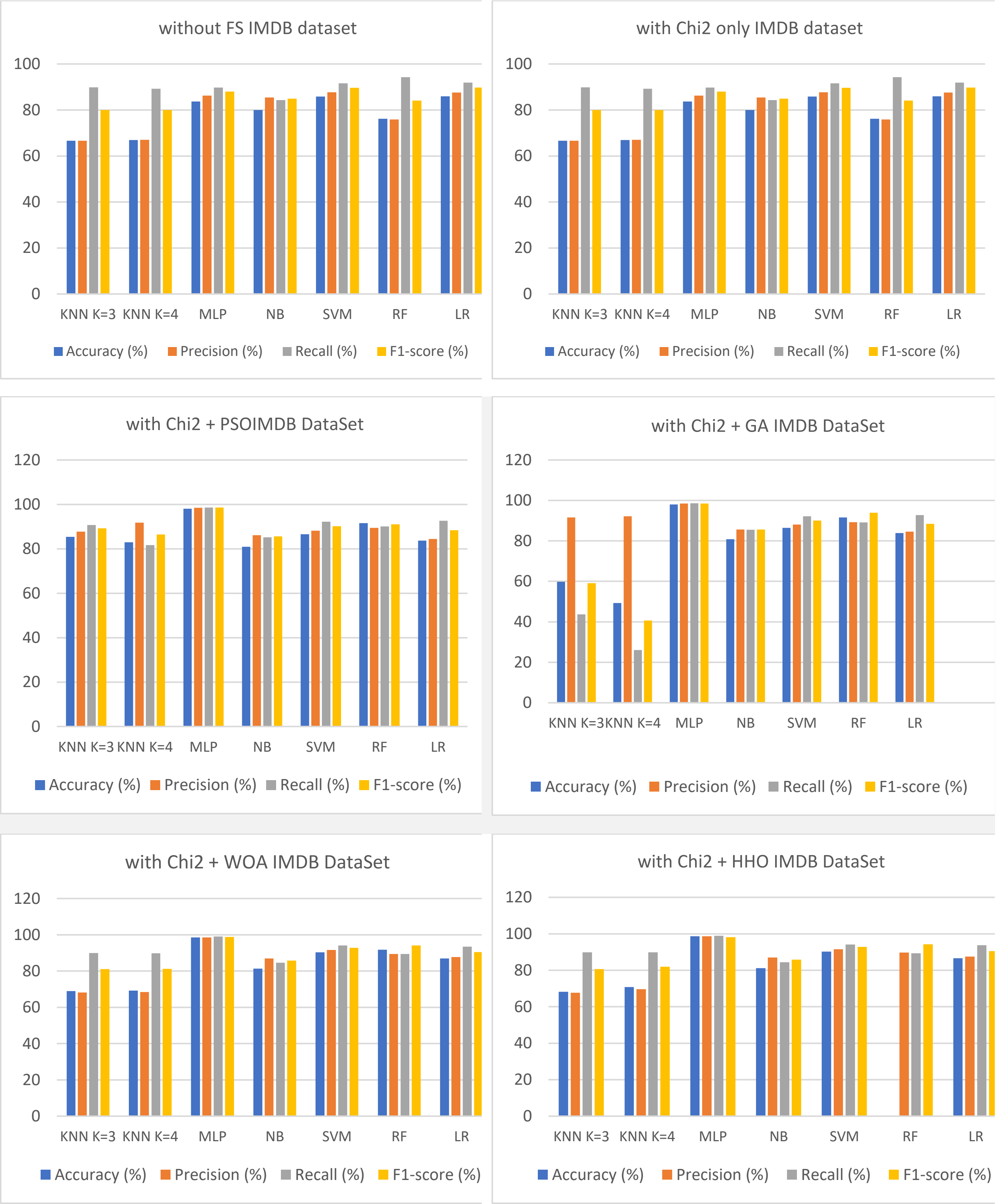

Figure 4: Accuracy, Precision, Recall, and F1-score of proposed model for all algorithms using IMDB dataset

Table 6: Results of Tweets us airline sentiment Dataset

Without FS Tweets us airline sentiment dataset					With Chi2 only Tweets us airline sentiment dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	63,05	67,22	21,25	20,01	KNN K=3	76,84	82,12	55,76	66,42
KNN K=4	62,36	65,15	28,92	26,3	KNN K=4	74,41	88,46	43,35	58,18
MLP	78,57	74,78	72,17	73,45	MLP	85,86	85,31	79,21	82,15
NB	81,42	81,00	71,56	75,99	NB	85,86	83,87	78,10	80,88
SVM	83,42	83,41	74,44	78,68	SVM	86,52	86,33	79,82	82,95
RF	79,67	84,27	62,08	71,50	RF	81,92	83,18	70,17	76,12
LR	84,06	84,85	74,50	79,34	LR	86,42	87,33	78,33	82,58
With Chi2+PSO Tweets us airline sentiment dataset					With Chi2+GA Tweets us airline sentiment dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	80,80	74,52	80,93	77,60	KNN K=3	87,07	89,31	77,83	83,17
KNN K=4	83,97	93,65	65,41	77,02	KNN K=4	83,17	91,96	64,69	75,95
MLP	92,16	92,99	87,52	90,18	MLP	91,55	92,93	85,98	89,32
NB	86,24	86,99	78,21	82,37	NB	85,56	86,42	76,94	81,41
SVM	87,73	89,21	79,77	84,22	SVM	86,61	88,05	77,99	82,72
RF	87,66	94,88	73,95	83,11	RF	88,00	94,55	75,11	83,72
LR	83,92	88,07	70,40	78,25	LR	83,47	88,06	69,12	77,53
With Chi2+WOA Tweets us airline sentiment dataset					With Chi2+HHO Tweets us airline sentiment dataset				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN K=3	86,00	91,08	73,06	81,08	KNN K=3	86,43	91,03	74,28	81,81
KNN K=4	81,35	94,41	58,03	71,88	KNN K=4	81,24	94,46	57,70	71,64
MLP	97,56	97,64	96,40	97,01	MLP	97,20	97,73	95,40	96,55
NB	86,57	85,96	80,43	83,10	NB	86,59	86,49	79,82	83,02
SVM	89,64	90,22	83,87	86,93	SVM	89,48	90,27	83,37	86,69
RF	89,34	96,58	76,77	85,55	RF	88,84	95,88	76,11	84,86
LR	85,52	89,03	73,83	80,73	LR	85,29	89,23	73,00	80,30

Working with the Tweets US Airline Sentiment dataset **-Table 6-** clearly shows how important feature selection is for improving model performance. When no reduction is applied, some models—especially KNN—perform poorly, with low recall and F1-scores. Just adding the Chi2 feature selection method already makes a noticeable difference, particularly for KNN, Naive Bayes, Random Forest, and MLP. But the real game-changer comes when Chi2 is combined with optimization algorithms like PSO, GA, WOA, and HHO. These combinations push models like MLP to achieve over 97% accuracy, and models like SVM, RF, and Logistic Regression also see significant improvements across all key metrics. PSO shows consistently strong results across the board, while WOA and HHO are especially effective with more complex models. In short, combining Chi2 with these optimization techniques not only boosts performance but also makes the models more reliable and better suited to handle real-world, messy data like tweets.

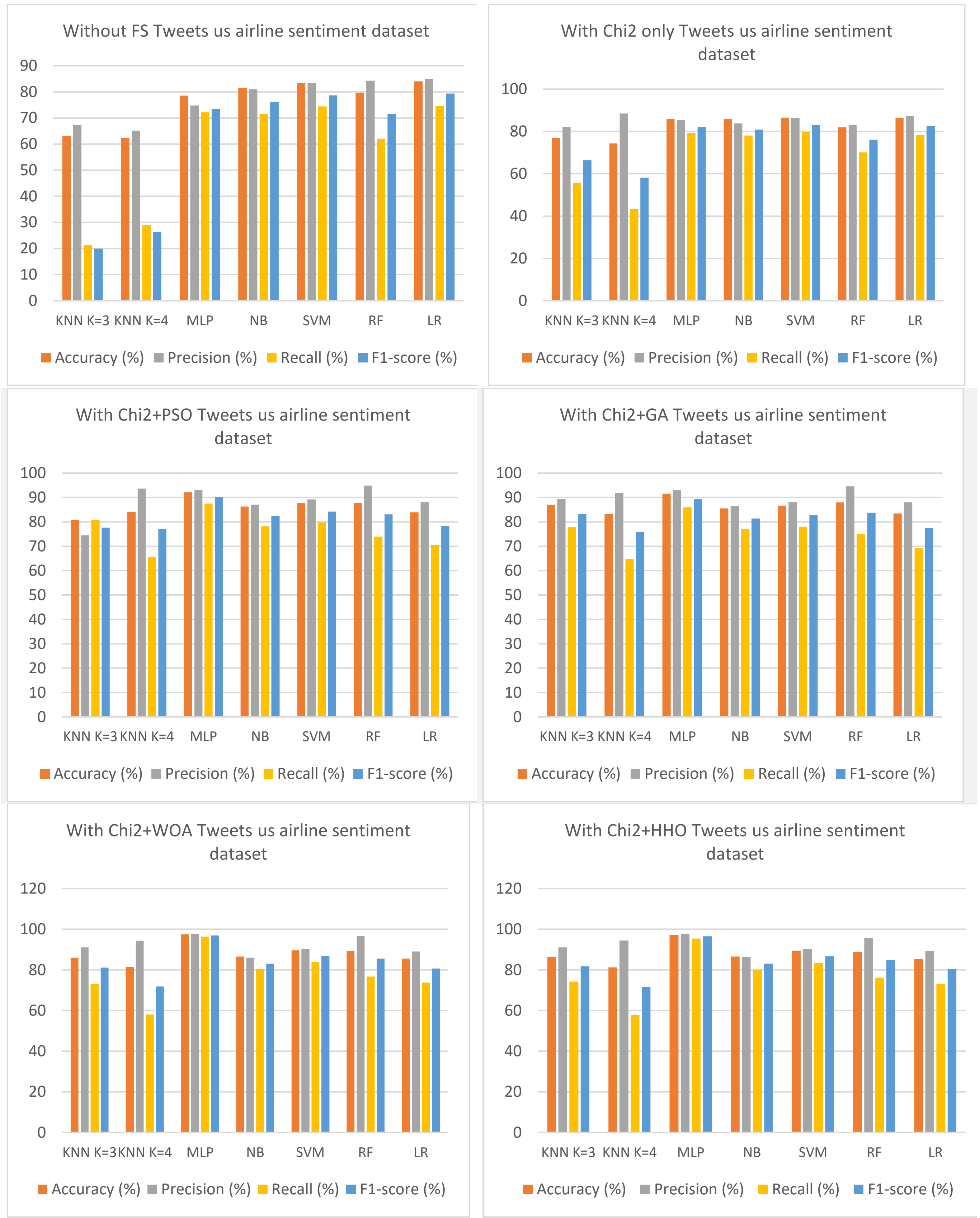

Figure 5: Accuracy, Precision, Recall, and F1-score of proposed model for all algorithms using Tweets us airline sentiment dataset

CONCLUSION

The results of this study highlight just how powerful hybrid feature selection methods can be for improving sentiment analysis on Twitter data. When no feature reduction is applied, many traditional models struggle—especially in terms of recall and F1-score—mainly because of the high dimensionality and noise typical of text-based data. Even using the Chi-square (Chi2) method on its own leads to noticeable performance improvements across several models. But the real breakthrough comes when Chi2 is combined with bio-inspired optimization algorithms like PSO, GA, WOA, and HHO. These hybrid techniques deliver significant gains in accuracy, with models like MLP reaching over 97%, and others such as SVM, Random Forest, and Logistic Regression showing steady improvements across all key performance metrics. PSO showed strong results across the board, while WOA and HHO stood out particularly on more complex models. All in all, these findings show that blending statistical and bio-inspired approaches makes models not only more accurate but also more reliable—offering a promising path forward for handling the challenges of sentiment analysis on noisy, real-world social media data.

REFRENCES

[1] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[2] Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2017). Sentiment Analysis: A Practical Guide. Springer.

[3] Pang, B., & Lee, L. (2008). "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135.

[4] Medhat, W., Hassan, A., & Korashy, H. (2014). "Sentiment Analysis Algorithms and Applications: A Survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113.

[5] Gao, L., Wang, P., Wang, J., & Wang, F. (2019). "A Review of Sentiment Analysis Methods Based on Deep Learning," IEEE Access, vol. 7, pp. 128279-128293.

[6] Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47.

[7] Guyon, I., & Elisseeff, A. (2003). "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

[8] Forman, G. (2003). "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research, vol. 3, pp. 1289-1305.

[9] Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2019). "A Survey on Evolutionary Computation Approaches to Feature Selection," IEEE Transactions on Evolutionary Computation, vol. 23, no. 3, pp. 421-441.

[10] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46-61.

[11] Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). "Harris Hawks Optimization: Algorithm and Applications," Future Generation Computer Systems, vol. 97, pp. 849-872.

[12] Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2020). "Hybrid Feature Selection Using Bio-Inspired Algorithms for Sentiment Analysis," Neurocomputing, vol. 400, pp. 220-232.

[13] Chandra, R., & Aggarwal, M. (2021). "Bio-Inspired Optimization for Feature Selection in Sentiment Analysis," Applied Soft Computing, vol. 101, 107042.

[14] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," Proceedings of the 11th International Semantic Web Conference (ISWC), 2012.

[15] C. N. dos Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," Proceedings of COLING 2014, pp. 69–78, 2014.

[16] A. Siddiqua, N. Asadullah, and S. M. A. Hossain, "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, vol. 145, no. 11, 2016.

[17] S. Silva, C. Ribeiro, and J. P. Carvalho, "An extensive analysis of feature selection techniques for Twitter sentiment classification," Expert Systems with Applications, vol. 42, no. 22, pp. 9293–9308, 2015.

[18] Chakraborty, I., Bhowmick, N., & Konar, A. (2016). "Feature selection using particle swarm optimization for sentiment classification." Applied Soft Computing, 46, 795–808.

[19] A. Bermingham and A. F. Smeaton, "Classifying sentiment in microblogs: Is brevity an advantage?" Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM), pp. 1833–1836, 2010.

[20] S. Dhuliawala, D. Kanojia, and P. Bhattacharyya, “SlangNet: A WordNet like resource for English Slang,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4329–4332.

[21] I. Aljarah, S. Sharieh, and H. Faris, "Feature selection using hybrid whale optimization algorithm for Twitter sentiment analysis," Procedia Computer Science, vol. 141, pp. 245–252, 2018.

[22] M. Habibi and M. Popescu, "A Hybrid Filter-Wrapper Approach for Feature Selection Applied to Twitter Sentiment Classification," Applied Soft Computing, vol. 77, pp. 661–670, 2019.

[23] M. Asghar, A. Khan, and F. Muhammad, "Feature Selection and Classification for Twitter Sentiment Analysis: A Hybrid Approach," Lecture Notes in Computer Science (LNCS), vol. 10573, pp. 528–540, 2017.

[24] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," CS224N Project Report, Stanford University, 2009.

[25] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and A. Potts, “Learning word vectors for sentiment analysis,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., 2011, pp. 142–150.

[26] K. L. Tan, C. P. Lee, K. S. M. Anbananthen and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," in *IEEE Access*, vol. 10, pp. 21517-21525, 2022.

[27] Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inform Technol Springer: 1–11

[28] Jianqiang Z, Xiaolin G (2017) Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 5:2870–2879

[29] Kennedy, J., & Eberhart, R. (1995). "Particle Swarm Optimization." *Proceedings of the IEEE International Conference on Neural Networks*, 1942–1948.

[30] Holland, J. H. (1975). "Adaptation in Natural and Artificial Systems." *University of Michigan Press*.

[31] Camacho, S., Vanneschi, L., & Castelli, A. (2020). "Genetic programming for Twitter sentiment analysis." Soft Computing, 24, 1867–1887.

[32] Bhatia MPS, Kumar A (2008) A primer on the web information retrieval paradigm. J Theoret Appl Inform Technol 4 (7).

[33] Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Elect Eng 40(1):16–28.

[34] Kumar, A., Jaiswal, A. Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. *Multimed Tools Appl* 78, 29529–29553 (2019).

[35] Moustafa, Maaskri, Mokhtar-Mostefaoui, Sid Ahmed, Hadj-Meghazi, Madani, & Goismi, Mohamed. (2024). Multi-Class Sentiment Analysis of COVID-19 Tweets by Machine Learning and Deep Learning Approaches. *Computación y Sistemas*, *28*(2), 507-516.

[36] Afriyani, S., Surono, S., & Solihin, I. M. (2024). Chi-Square Feature Selection with Pseudo-Labelling in Natural Language Processing. *JTAM (Jurnal Teori dan Aplikasi Matematika)*, *8*(3), 896-909.