Neural Computing and Applications (2021) 33:15091–15118
https://doi.org/10.1007/s00521-021-06406-8

REVIEW ARTICLE

A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities

Esther Omolara Abiodun[1,3] · Abdulatif Alabdulatif[2] · Oludare Isaac Abiodun[1,3] · Moatsum Alawida[1,4] · Abdullah Alabdulatif[5] · Rami S. Alkhawaldeh[6]

Received: 7 April 2021 / Accepted: 31 July 2021 / Published online: 13 August 2021

Abstract

Specialized data preparation techniques, ranging from data cleaning, outlier detection, missing value imputation, feature selection (FS), amongst others, are procedures required to get the most out of data and, consequently, get the optimal performance of predictive models for classification tasks. FS is a vital and indispensable technique that enables the model to perform faster, eliminate noisy data, remove redundancy, reduce overfitting, improve precision and increase generalization on testing data. While conventional FS techniques have been leveraged for classification tasks in the past few decades, they fail to optimally reduce the high dimensionality of the feature space of texts, thus breeding inefficient predictive models. Emerging technologies such as the metaheuristics and hyper-heuristics optimization methods provide a new paradigm for FS due to their efficiency in improving the accuracy of classification, computational demands, storage, as well as functioning seamlessly in solving complex optimization problems with less time. However, little details are known on best practices for case-to-case usage of emerging FS methods. The literature continues to be engulfed with clear and unclear findings in leveraging effective methods, which, if not performed accurately, alters precision, real-world-use feasibility, and the predictive model's overall performance. This paper reviews the present state of FS with respect to metaheuristics and hyper-heuristic methods. Through a systematic literature review of over 200 articles, we set out the most recent findings and trends to enlighten analysts, practitioners and researchers in the field of data analytics seeking clarity in understanding and implementing effective FS optimization methods for improved text classification tasks.

Keywords Feature selection · Hyper-heuristics · Metaheuristic algorithm · Optimization · Text classification

1 Introduction

In the last few decades, the world has witnessed the proliferation of the Internet amongst people, organizations and governments [1, 2]. Modern architectures, such as the Internet of things (IoT), Internet of medical things (IoMT), industrial Internet of things (IIoT), Internet of flying things (IoFT), amongst others, unlatch incredible opportunities for the realization of intelligent living and well-being of humanity [3, 4]. Consequently, a massive amount of digital data is generated on a daily basis. The generated data, which are in the form of texts, numbers, audios, videos, tapes, graphs, images and so forth, are extensions of knowledge. In this regard, data from diverse spheres of life such as health, agriculture, transportation, finance,

✉ Esther Omolara Abiodun
aeomolara@student.usm.my

1 School of Computer Sciences, Universiti Sains Malaysia, George Town, Malaysia

2 Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia

3 Department of Computer Sciences, University of Abuja, Abuja, Nigeria

4 Department of Computer Sciences, Abu Dhabi University, Abu Dhabi, UAE

5 Computer Department, College of Sciences and Arts, Qassim University, P.O. Box 53, Al-Rass, Saudi Arabia

6 Department of Computer Information Systems, The University of Jordan, Aqaba 77110, Jordan

education, sport, amongst others, can be categorized and subsequently leveraged for knowledge, insights and predictions.

A substantial amount of the knowledge available these days are stored as text [5]. A recent analysis from Forbes reported that about 2.5 quintillion bytes of data are generated daily [6]. The report also showed that a large portion of the generated data was in textual form. For instance, Facebook records over 20 billion messages in textual, pictorial, audio and video forms [7, 8]. Likewise, over 15 billion tweets are exchanged on Twitter pages on a monthly basis [9]. In addition, the English Wikipedia contains about 6,272,058 articles, and it averages around 604 new articles each day [10].

Data mining, acoustics, pattern recognition and text analysis specifically aim to recognize peculiarities within data by simulating and extracting data content. It leverages a number of methods from the domain of artificial intelligence, statistics and so forth to classify texts in documents, news, web pages and others from the field that characterize the problem that is to be resolved. Text classification consists of preparing the data by transforming the raw data into a suitable form for modelling. It is a general consensus in the field of data mining that your model is only as good as your data. Hence, data preparation techniques are an essential requirement to get the most out of data and in turn, generate a predictive model with optimal performance. Raw data cannot be utilized directly due to certain issues. For instance, implementations may require that data be numeric; raw data may contain errors, algorithms may mandate explicit requirements, columns or segments may be repetitive, redundant, irrelevant or insignificant.

Text mining (inclusive of text data mining) techniques allow the discovery of high-quality information from texts. It encompasses data cleaning, feature selection, data transforms, feature engineering, dimensionality reduction and so forth. Every one of these tasks is an entire field of study with specialized algorithms. However, this study focuses specifically on feature selection. The basic steps of the text classification process are shown in Fig. 1.

Figure 1 depicts the set of processes in the text data mining process. It begins with text processing which is a data preparation process encompassing tokenization, word normalization, stop word removal, filtering, amongst others. This is followed by the feature extraction phase and then the feature selection phase before the interpretation of the model.

Feature selection (FS) seeks to enhance classification efficiency by selecting only a tiny subset of appropriate features from the initial wide range of features. FS attempts to find an optimal set of features by removing redundant and unimportant features from the dataset. The removal of irrelevant and redundant features yields a good text representation, a decreased data dimensionality, accelerates the learning cycle of the model, and boosts the performance of the predictive model. Hence, the advantage of feature selection ranges from minimizing overfitting, reducing data dimensionality, improving accuracy, eliminating irrelevant data, expediting training to improve insights and elucidating the intricacies within the data, amongst many other advantages.

The three main methods of feature selection for text classification are namely filter-based, wrapper-based and embedded. Each method of FS has its merits and demerits. Recent years have seen the progression of research towards combining two or more methods to produce the hybrid-based feature selection method for better text classification.

The convoluted and cumbersome nature of the entirety of most real-world problems requires an ample solution space due to interdependencies and nonlinear requirements amongst attributes [11]. Thus, the conventional-based feature selection techniques are unable to handle such problems. For instance, the filter-based methods have critical issues ranging from them being unable to increase consumption time, deliver satisfactory performance, complexity and others. These challenges and more have mandated researchers to explore diverse other methods of obtaining better performing options during the classification task. Hence, the pursuit of better techniques with optimal performance has led to the discovery of metaheuristic-based feature selection methods for text classification.

Metaheuristic-based algorithms have proven their suitability in diverse areas due to their delivery of practical solutions in considerable time and their specificity in overcoming the curse of dimensionality by optimizing the performance of classification, mitigating high use of computational resources, storage and the number of features. Examples of metaheuristic algorithms include ant colony optimization [12], genetic algorithms [13], memetic algorithm [14], particle swarm optimization [15], evolutionary-based algorithm [16], grey wolf optimizer [17], firefly [18], binary Jaya [19], dragonfly algorithm [20, 21] and so on.

This study focuses on metaheuristic-based feature selection algorithms for text classification due to their favourable characteristics of performing better than traditional-based feature selection methods. This review is urgently required because of the lack of accurate information on metaheuristic-based feature selection methods, which currently affects the practice, accuracy and general performance of most predictive models utilized for text classification in different domains.

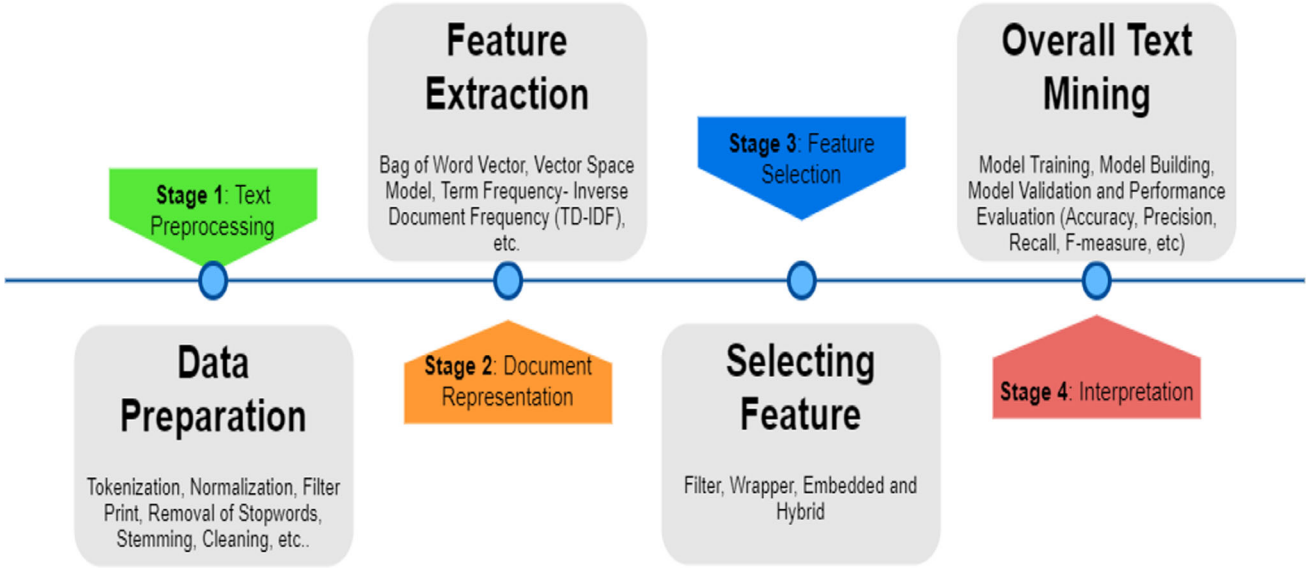

Fig. 1 Basic steps for text classification

1.1 Contributions

Various overviews of feature selection are available in the literature. For instance, Chandrashekar & Sahin [22] provided a general introduction to feature selection methods and classified them into the filter, wrapper and embedded. Pereira et al. [23] carried out a comprehensive survey and novel categorization of the feature selection techniques by focusing on multi-label classification. Notwithstanding, these reviews did not consider how metaheuristic algorithms affect or influence the accuracy of text classification and the analysis of the different methods to handle the high dimensionality of the feature space.

To the best of our knowledge, after a thorough scrutiny of the available literature, no work expounded on the emerging metaheuristic and hyper-heuristic optimizations for feature selection. Moreover, few articles gave detailed insights into their present state and prospects, alongside how flawed feature selection processes impact the practicality of the predictive model for real-world use cases. This paper serves such a purpose. Thus, the significant contributions of this research are as follows:

1.1.1 Encyclopaedic knowledge of feature selection

This review will serve as a concise encyclopaedia to analysts, practitioners, researchers and stakeholders in the field of data analytics seeking clarity in understanding the basic and advanced techniques of feature selection. It serves as an all-encompassing referential manual and guide for selecting effective and efficient feature selection optimization methods for optimal development of predictive model for text classification. Likewise, it could also serve as a fundamental framework to guide newcomers and interested researchers in the field.

1.1.2 Up-to-date overview

An updated overview of the current methods of feature selection is discussed in this study. It is an extended effort that can assist prospective researchers to immediately understand some essential concepts and know the keyword in the process of feature selection. The knowledge of the concept and keyword will help to save time and address any complexity of the feature selection process by guiding the potential researchers in designing remarkable fail-proof frameworks for optimizing their algorithms.

1.1.3 Extensive resources

This study examines the application of metaheuristics for feature selection. We investigate and assemble many resources on metaheuristic techniques that handle feature selection, including state-of-the-art models, real-world use-cases and characteristics of the benchmark datasets. This study briefly serves as a hands-on guide for understanding and generating peculiar feature selection models for categorizing, characterizing and modelling real-life practical scenarios.

Similarly, the study will serve as a handbook for discovering the suitable statistical and modelling approaches for feature selection, its significance and choosing the practical techniques to leverage for various variable types.

1.1.4 Open issues and future insights

An in-depth investigation, exploration and discussion on current trends of feature selection patterns, limitations and prospective future research directions are discussed. Such exploration serves as an apogee for the contributions and limitations of the reviewed studies to elucidate novel practices that could further advance this field.

The structure of the paper is mapped out as follows: (II) the feature selection process, which explains the concept of feature selection, (III) the review methodology section provides full details on how the papers were selected, (IV) the existing literature on the metaheuristic-based algorithm presents comprehensive details on the state-of-the-art of the metaheuristic-based methods, (V) research gaps were provided in this section, (VI) lessons learned during the review were discussed in this section, (VII) other issues and possible solutions contain details of other relevant information in the feature selection process, (VIII) future directions give details of potential opportunities (IX) and conclusion.

2 Feature selection

Feature selection is an essential data preparation technique performed to characterize the most relevant, pertinent and significant feature space. It involves selecting the subset of the most distinct and relevant feature from a large group of features to represent a record in a dataset for predictive modelling [24]. It is an aspect of feature engineering where the attribute or item of a dataset is utilized to reduce the dimensionality of the problem to be addressed and thus, facilitate the phase of the classification process. The primary motivation of the feature selection task is dimension minimization in a huge multi-dimensional dataset. The innovation of feature selection is a major step of successful knowledge discovery in a problem with a large number of features.

The main challenge of feature selection stems from picking the smallest number of features from the primary dataset, which occasionally consists of a large number of features. Finding specific relationships and arriving at a conclusion when dealing with a large dataset is quite difficult because some features are so related to the problem at hand while some others are not related. If all the features were selected, it would affect the selection outcome. Therefore, to find the best solution, it is essential to select the features that are most related only to the given problem. Additionally, any one of the features that can affect the outcome, which will lead to inaccurate results or that may be time-consuming in the analysis process, should be avoided. The ideology of minimizing the attributes in the large dataset during feature selection is represented in Fig. 2.

Figure 2 depicts the process where one can manually or automatically select those features from the original dataset which contribute most to the prediction variable or output in which one has an interest. Having irrelevant features in data can decrease the accuracy of the models and make a model learn based on irrelevant features. Thus, from the original dataset, a subset of data is created to eliminate irrelevant features.

In the feature selection process, attribute elimination can help in knowing the size of data, reducing computation time and requirement, minimizing dimensionality and improving the performance predictor. In addition, the selection of the features helps the predictive models to detect hidden intricacies that can improve the performance of the specific domain in view. For example, in the covid-19 control model, there is a need for early detection of covid-19, especially due to the lack of a widely known cure [25, 26]. The significant features that will be useful in the prediction are attributes encompassing major details of the patient's symptoms, such as if the person is having shortness of breath, fever, headache, sore throat, cough, muscle pain and fatigue. Personal details containing features like the height, weight of the person, phone number, residential address, etc., may be irrelevant for the prediction. Thus, such data clusters will not be included at the feature selection phase of developing the disease detecting model. Consequently, the model can be used for early discovery and prevention of the further spread of the covid-19 disease. Therefore, the purpose of the feature selection process is to reduce the number of features drastically. However, the reduction needs not jeopardize the accuracy of the model. Therefore, the success of the selection process heavily relies on two critical factors, increasing the rate of accuracy and minimizing the number of attributes [27].

The literature classifies the feature selection process into four, namely filter, wrapper, embedded and hybrid methods. An overview of the FS methods is given succinctly in the following subsection. The classification of feature selection methods is represented in Fig. 3.

2.1 Filter-based method

The filter approach applies an evaluation function to each element, and subset selection is performed depending on the score achieved. It evaluates features according to heuristics based on the general characterization of the data [28]. Statistical analysis is performed over the feature space via ranking each feature of the dataset based on some standard univariate metrics and then selecting the highest ranking features. Some of the metrics include:

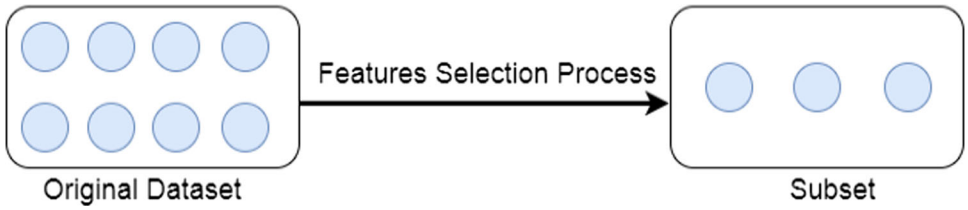

Fig. 2 Feature selection concept

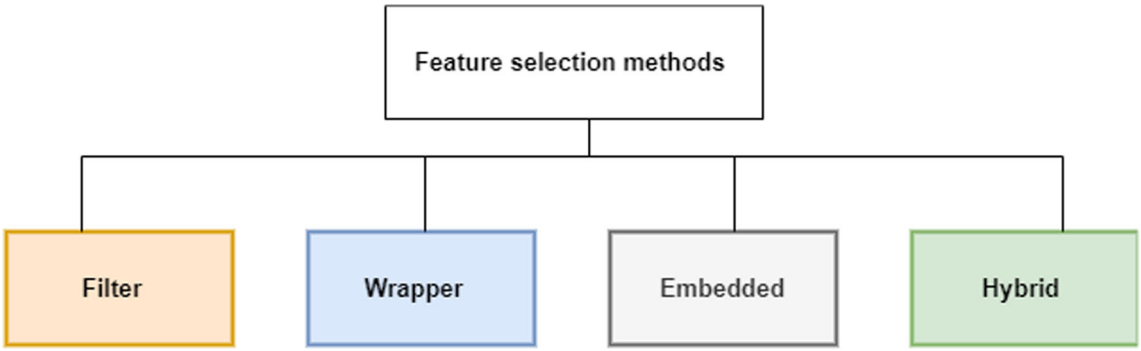

Fig. 3 Classification of feature selection methods

Correlation coefficients: This metric eliminates duplicated features.

Information gain: It accesses the independent parameter by predicting the target parameter.

Chi-square: It tests the independence utilized in deciding the dependency of two variables.

Variance: It eliminates constant features and other quasi constant features.

The metrics are disintegrated into several specific measures like Welch T-test [29, 30], Fisher score [31], Kendall correlation [32, 33], feature similarity [34], Pearson correlation [22, 35], correlation [36, 37], amongst others. The filter-based techniques could be used to select the best feature by using specific filter criteria or selecting independent features that have a high correlation with the target variable, low correlation with other independent variables and reciprocated information of the independent variable.

Compared to the wrapper-based method, especially during their application to large datasets, the filter-based methods have the advantage of performing faster in computation with minute computational time. Likewise, they are robust to overfitting and model agnostic as they depend entirely on the features in the dataset sample. The filter-based methods also use relations between one input attribute and the output attribute and search locally for attributes that permit good local discrimination [38]. Nevertheless, redundant features might not be filtered as they work more with discrete classification problems. Such problems are being resolved in the literature by attaching other metrics. For instance, Hall [28] presented a fast-correlation-based FS method suitable for addressing discrete and continuous classification problems.

2.2 Wrapper-based method

The wrapper-based FS method utilizes the learning algorithm itself to assess the usefulness of features. The wrapper method creates an interaction between the classification algorithm and the search subset. It implements a subroutine, which acts as a statistical resampling technique (for instance, cross-validation) utilizing the actual target learning algorithm to estimate the accuracy of feature subsets. The wrapper approach has demonstrated its superiority in classification tasks, as they perform well when solving the "real world" problem by optimizing the classifier performance. However, it is prolonged during execution as the learning algorithm has to be called repeatedly. Compared to the filter method, they are computationally more tedious due to the repetitive learning steps and cross-validation. Wrapper FS methods do not scale well to enormous datasets containing many features. Although, the results can be more accurate than in the previous filter-based method [39]. Nevertheless, this can lead to a longer time to get results than the previous method since it requires that the classifier be used severally. Examples of the "wrapper method" include genetic algorithms, sequential algorithms and recursive feature elimination [40]. A particular case of sequential feature selection is a greedy search algorithm that could locate the "optimal" feature subset by iteratively selecting features based on the performance of a classifier. It starts with a null feature subset and adds one feature one after the other in each

round. One feature can be selected from the pool of all features that are not in the original subset, but the results become the best performance classifier if added. The general wrapper approach is demonstrated in Fig. 4.

As presented in Fig. 4, the wrapper method utilizes a predefined classifier to explore a subset of features. It then applies the classifier to measure the selected subset of features. The selection and measuring of subsets of features continue till the desired criterion of quality is achieved.

2.3 Embedded method

The embedded feature selection methods are implemented using algorithms with their own built-in feature selection methods. It is similar to the wrapper method in which the same classifier is employed in selecting attributes at the evaluation phase. However, using the classifier in the embedded method is achieved at a less computational cost than the wrapper method [22]. Popular examples of such methods are decision trees, RIDGE, least absolute shrinkage and selection operator (LASSO) and regression with inbuilt penalization functions to reduce overfitting. At the same time, LASSO regression is a regularization technique used over regression methods for a more accurate prediction. RIDGE regression is a technique for analysing multiple regression data that suffer from multicollinearity (correlations between predictor variables). For instance, to develop a parsimonious model, ridge regression is employed as a strategy to determine if the number of predictor variables in a set exceeds the number of observations or when a dataset has multicollinearity.

LASSO ($L1$) regression for generalized linear models might be understood as adding a penalty against complexity to reduce the degree of variance or overfitting of a model by putting additional bias. That is, adding a penalty term directly to the cost function,

Regularized cost = regularization penalty + cost.

In $L1$ regularisation, the penalty term is,

$L1 : \lambda \sum k_i w_i = \lambda w_{1,} \ldots (1)$where w is a k-dimensional feature vector. By adding the $L1$ term, the objective function now becomes the minimization of the regularized cost. Since the penalty term grows with the value of the weight parameters that is λ just a free parameter to fine-tune the regularisation strength, one can induce sparsity through this $L1$ vector norm, which may be considered as an intrinsic way of feature selection which comprises of the model training step. Meanwhile, the process of an embedded method is illustrated in Fig. 5.

2.4 Hybrid method

The hybrid technique utilizes more than one strategy for selecting a feature to create subsets. It combines multiple approaches to obtain the best possible feature subset rather than using an independent method. In the hybrid approach,

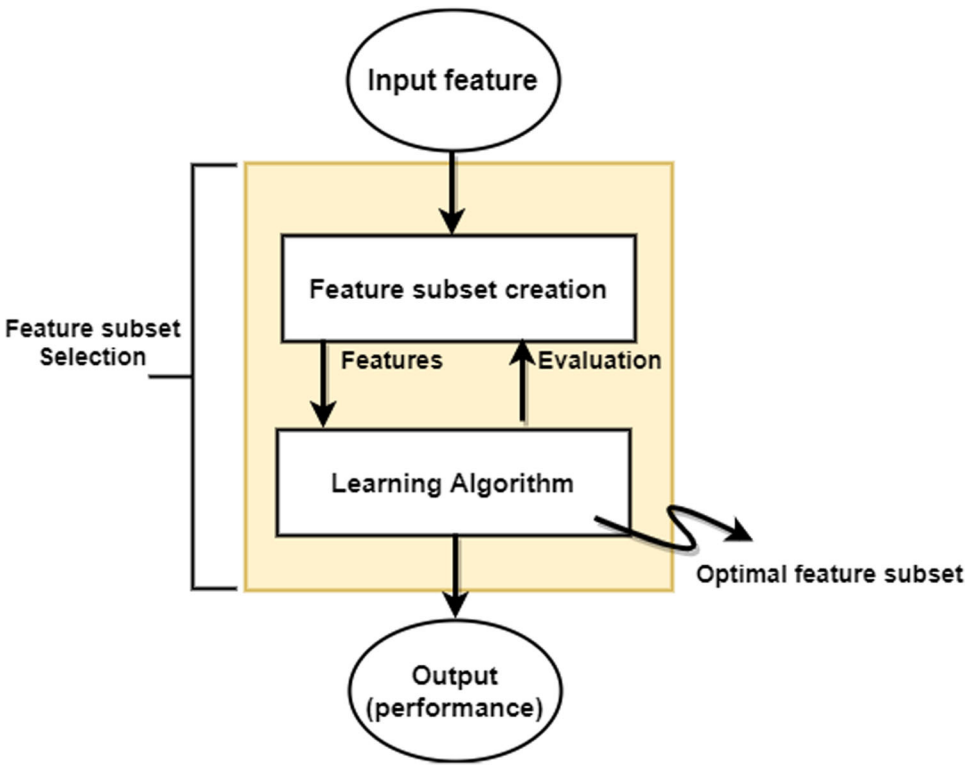

Fig. 4 The process of wrapper model

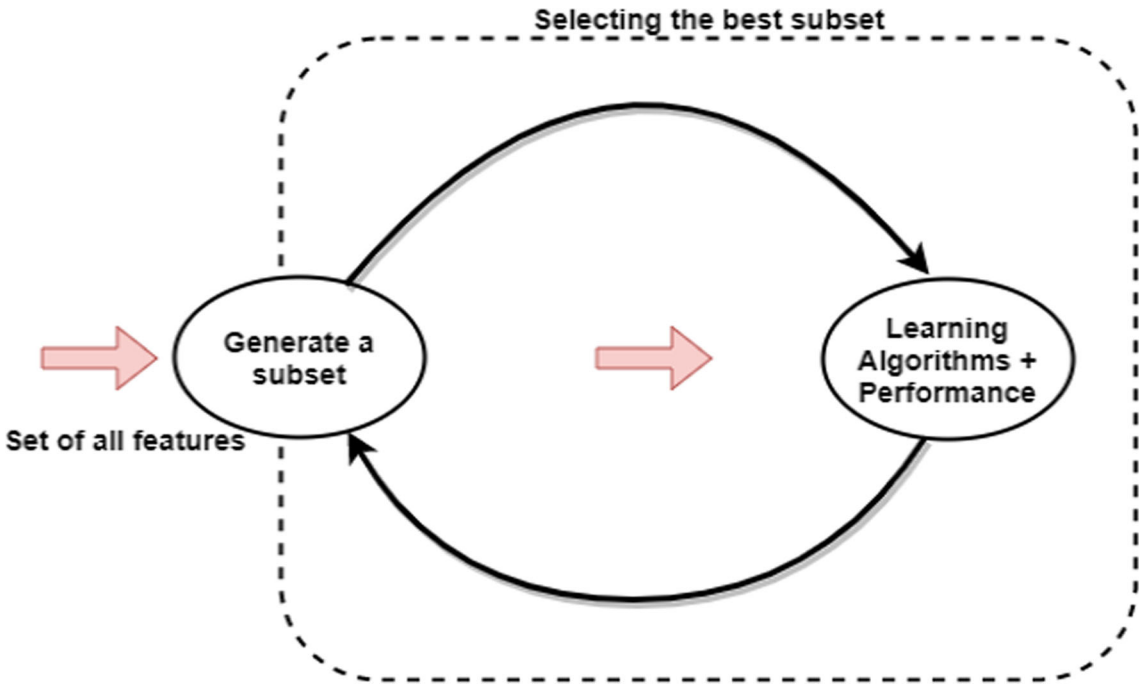

Fig. 5 The process of an Embedded Model

two methods can be combined logically, for instance the wrapper and filter method. It begins with the filter method being used to create a subset of features, followed by the wrapper method being used to select features from the subset [41]. The hybrid method can take advantage of the wrapper and filter methods by exploiting their different evaluation benchmark in different search phases. Then achieve a relative comparable accuracy to the wrapper method and also comparable efficiency to the filter method. It first incorporates the statistical criteria, as the filter method does, to select various candidate features subsets with a specific cardinality. Then, it selects the subset with the highest classification accuracy, just as the wrapper does.

Combining these methods depends on each person performing the feature selection, given that one has many methods in the toolbox. For example, the modeller may begin by performing the filter method (such as removing constant, duplicated features and quasi-constant). The next step involves using the wrapper method to select the best feature subset from the previous step. The hybrid method builds on the intuition of creating an effective and efficient model by combining weaker methods, thus, the term hybrid. The hybrid methods can perform both feature selection and model training concurrently. A high accuracy and performance, optimal computational complexity, robust and flexible models are some of the benefits enjoyed from the hybrid methods. The hybrid methods can combine filter and wrapper methods of feature selection simultaneously as depicted in Fig. 6.

Although the hybrid methods often offer a superb way of combining weak feature selection methods to achieve better selection variables, the drawback is that it can be expensive and time-consuming when combining different methods. Some of the merits and demerits of the selection of features methods are shown in Table 1.

3 Review methodology

This review overviews and reports the current state of the metaheuristic and hyper-heuristic optimization methods. It investigates and examines the literature on current feature selection methods, metaheuristic, hyper-heuristic optimization methods and much more. Detailed studies of related works from the literature were reviewed to achieve the objectives of the review. Reviewed works of the literature were extracted from the vast resources in well-established and reputable databases containing published articles from popular journals, conference papers and proceedings, books, edited volumes, thesis, symposiums, preprints, grey literature, government and organization publications, magazines and lecture notes amongst others.

The relevant works in the literature were identified by querying related search terms such as “Feature Selection”, “Hyper-heuristics”, “Metaheuristic Algorithm”, “Optimization”, “Text Classification”, “Data Mining” and

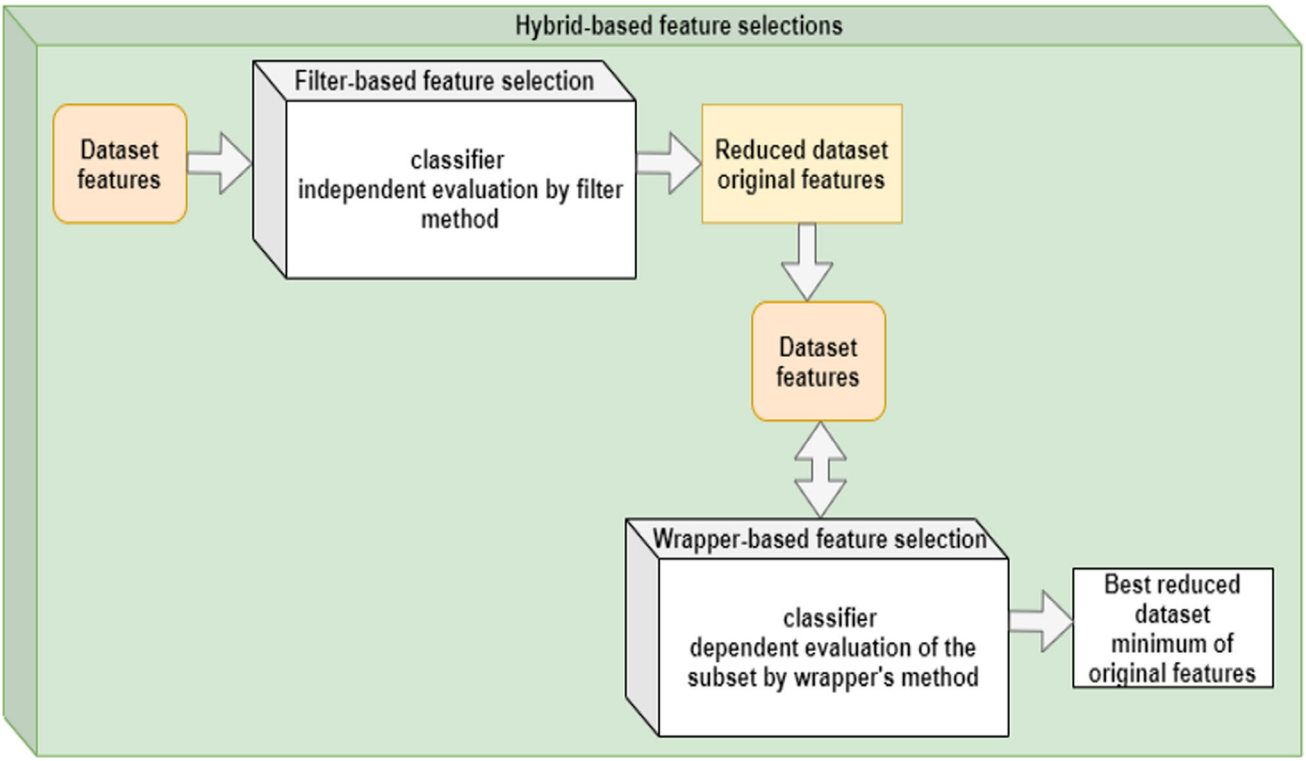

Fig. 6 A hybrid method combining both filter and wrapper methods of feature selection

"Text Data Mining". Finally, the search keywords used include: "Problems and Solutions in Meta-Heuristic Algorithm", "Problems and Solutions in Hyper-Heuristic Algorithm", "Future Prospects in Meta-Heuristic Algorithm", "Future Prospects in Hyper-Heuristic Algorithm", "Optimization Methods" and "Classification Tasks". The returned results were downloaded, read and relevant papers were collated for the final analysis. The scholarly databases queried for the literature are:

IEEE Xplore,
Science Direct,
ACM Digital Library,
Scopus,
Elsevier,
Springer,
EBSCO Host,
Taylor and Francis,
Research Gate,
And Google Scholar.

The major synthesis of the research concentrated on recent work between the year 2015–2021. Thus for the ***inclusion criteria,*** we considered:

Studies published from the year 2015–2021 which are related to metaheuristic-based text feature selection.

Studies that are published strictly in peer-reviewed journals.

For the ***exclusion criteria,*** we considered:
Studies published in unknown journals.
Studies with redundant information. For instance, we selected the extensive study in a case whereby the same study is published in a conference and a journal.

Overall, 200 papers were used for the review. The summary of the articles processed in the review is clearly explained in Table 2.

Thus, Table 2 shows the summary of the number of articles processed in the review.

4 Metaheuristic-based algorithms

The intricacies of FS problems emanate from selecting the most relevant set of features from an abundance of large possible subsets. FS introduces a combinatorial problem that is not easily solved using traditional feature selection and optimization techniques. Thus, heuristics-based algorithms found their way into the picture and have become more established in the literature in the quest of finding better solutions for complex challenges.

Metaheuristic-based algorithms are leveraged for addressing numerous kinds of optimization problems utilizing self-learning operators configured with actors to

Table 1 Merits and demerits of feature selection methods

Filter-based methods		Wrapper-based methods	Wrapper-based methods
Merits	Demerits	Merits	Demerits
Operate independently of any learning algorithm	No interaction with classification model for feature selection	Wrapper methods have good generalization than filter methods	The approach is slow as the algorithm has to be called repeatedly
Undesirable features are filtered out of the data before induction commences	Filter methods may miss features which may be independently irrelevant but are very useful influencer when combined with other features	They can be "wrapped" around any continuous or discrete class learner	As the number of input features increases, it becomes computationally costly
Lower risk of overfitting	Most existing filter algorithms perform well only while handling discrete classification problems	Wrapper-based methods retain the feature set that yields the best accuracy	They do not scale well to large datasets consisting of numerous features
Filter methods are model agnostic	It presents the challenge of finding the threshold point for ranking to choose only the required features while excluding noise	Models feature dependencies between each of the input features	It is not model agnostic
Computationally cheaper compared to the wrapper and embedded methods	It is less accurate when compared to other advanced feature selection methods like the hybrid	Dependent on the model selected	The risk of overfitting is high
It is computationally very fast		Interact with the classifier for feature selection	Classifier dependent selection
They are scalable as they are based on different statistical methods		More detailed search of feature set space	Longer running time
They consider relations between one input attribute and the output attribute and also search locally for attributes that allow good local discrimination			No guarantee of optimality of the solution if predicted with another
Filter methods rely entirely on the features in the dataset			computationally infeasible with an increasing number of features
Filter methods have the ability of good generalization			
Quickly scale to high-dimensional datasets			

Embedded methods		Hybrid methods	Hybrid methods
Merits	Demerits	Merits	Demerits
It outperforms the filter method in generalization error with an increased number of data points	Considers the dependence among features	Less prone to overfitting problems	The technique of developing hybrid-based methods may be quite gruesome as it requires incorporating more than one method
Provides feature importance for better accuracy	Classifier dependent selection	It can find the feature subset for the algorithm being trained	
Less prone to overfitting problems	Identification of a small set of features may be problematic	They are also faster in achieving optimal solutions	
It takes into consideration the interaction of features		It takes into cognizant the interaction of features	
They are also faster compared to the filter methods			
They can find the feature subset for the algorithm being trained			
Less computational intensive compared to the wrapper			

effectively investigate and manoeuvre probable solutions with an expectancy of arriving at the best solution [42]. They are nature-inspired algorithms based on scientific principles from biology, ethology, mathematic, physics, amongst others. Additionally, they are identified as high-level problem algorithmic schemes that provide a beehive of strategies, rules or guidelines to design heuristic optimization algorithms [43].

Heuristics are strategies such as the rules of thumb, common sense and error. Metaheuristics are general ideas, techniques or methods that are not particular to a singular problem [44, 45]. Metaheuristics are estimating paradigms in which each algorithm has a different historical

Table 2 Summary of the number of articles processed in the review

Indexer	Results	Profiteered	Relevant
IEEE Xplore	125	20	70
Science direct	42	17	19
ACM digital library	33	15	17
Scopus	30	18	10
Elsevier	28	9	12
Springer link	28	10	13
EBSCO host	26	9	14
Taylor and Francis	20	7	9
World of science (WoS)	25	8	11
Research gate	20	6	8
Google scholar	29	8	10
Others	28	8	7
Total	434	135	200

background [46, 47]. Likewise, they are seen as a set of algorithmic concepts utilized for defining heuristic techniques that can be applied to diverse optimization problems with slight modifications to adapt them to specific problems [48, 49].

In recent times, metaheuristics have been successfully utilized for addressing classification problems. Metaheuristics are introduced into feature selection in various fields on account of their excellent global search capability and performance. They have been applied for many real-world optimization challenges, including load balancing in telecommunication networks and flight schedules, economic load dispatch problem [50], gene selection in cancer classification in the medical domain [51], amongst others.

The established literature categorizes the metaheuristic-based algorithms into a population and local search algorithm [42]. The population-based algorithms examine a number of search space regions simultaneously and enhance them iteratively to attain the ideal solution. Examples of population-based algorithms are genetic algorithm, ant lion optimizer, firefly, bat algorithm, competitive swarm optimizer, whale optimization algorithm, differential evolution, crow search [52], etc.

The local search-based algorithms consider one solution (referred to as the initial solution) at a time. It is remodelled persistently by utilizing an operator which allows visiting relatively close values until a peak local value is obtained. It locates the local optima by exhaustively exploring certain regions of the initial solution. Notwithstanding, the inability of exploring multi-search space regions simultaneously is a limitation. Thus, some methodologies are employed to empower the local search-based approach. Instances of such techniques leveraged in the search are tabu search [53], stochastic local search method [53], iterated local search [54], variable neighbourhood search [55], GRASP [56], etc.

An extensive treatment of various metaheuristic algorithms' references can be found in the work [42, 52]. A detailed description of the state of art is given in the following subsection.

4.1 The state of the art: metaheuristics methods for text classification

Nowadays, feature selection methods based on metaheuristics are increasingly studied and applied due to the importance and necessity of feature selection. Metaheuristics methods of feature selection are majorly classified into swarm intelligence, evolutionary-based and trajectory-based algorithm. A thorough synthesis and discussion of each algorithm and their classes of sub-methods are given as follows:

4.1.1 Swarm intelligence (SI)

Swarm intelligence (SI) is a population-based stochastic optimization technique that emerged as a family of nature-inspired algorithms. It describes the aggregate behaviour of decentralized, coordinated and self-organized frameworks that can move rapidly in a planned way. The framework comprises a population of simple agents that can directly or indirectly communicate locally by acting on their local environment [57]. Some examples are ant colonies, bee colonies, animal herding, birds flocking, fish schooling, hawks hunting, bacterial growth and microbial intelligence [58]. Generally, they provide robust solutions to different complex problems.

Examples of the SI-based metaheuristic method for feature selection are particle swarm optimization (PSO), artificial bee colony optimization (ABC), ant colony optimization (ACO), bat algorithm (BA), gravitational search algorithm (GSA), firefly algorithm (FA), cuckoo optimization algorithm (COA), salp swarm algorithm (SSA), whale optimization algorithm (WOA), grey wolf optimization (GWO), amongst others. Recent researches on each method are given in the subsequent paragraphs.

The PSO-based algorithm is motivated by the social behaviour of birds and fish. *In the PSO-based method*, [59] put forward a Hamming distance-based binary PSO(HDBPSO) algorithm to reduce data dimensions. The technique selects the relevant features by using hamming distance to update the velocity of particles in a binary PSO search procedure. [60] proposed an improved multi-objective PSO method to enhance the searchability of the PSO-based approach based on the introduction of two new operators. [61] presented an integration of correlation FS with a modified binary PSO algorithm to classify cancer

and select genes. [62] proposed a cross-bred PSO-based FS to enhance the accuracy of laser-induced breakdown spectroscopy analysis. Other notable works in the literature based on the PSO-based approach can be found in [63–68].

The ABC-based method is inspired by the intelligent behaviour of the simulating food search behaviour of bee groups/populations. *In the ABC-based method* [69], put forward a hybrid of ABC and integrated it with the ACO to produce a high performing model. A two-archive multi-objective ABC algorithm was presented by [70]. An increase in the accuracy and a lesser computational complexity was attained by integrating a multi-objective optimization algorithm with a sample reduction technique using ABC [71]. [72] presented a variant of ABC called a multi-hive artificial bee colony for high-dimensional symbolic regression with feature selection. Grover and Chawla used an intelligent strategy to improve the ABC algorithm [73]. Other notable contributions using the ABC approach in the literature can be found in [74–77].

The ACO-based method is motivated by the behaviour of ants searching the shortest path to get food in between the nest and the food source and their adaptation to natural changes. *In the ACO-based method,* a clustered graph is utilized to represent the FS problem based on the ACO and social network analysis [78]. An unsupervised probabilistic FS that searches for the optimal feature subset in an iterative schedule by leveraging the similarity between the features using ACO was presented by [79]. [80] improved the classification accuracy of imbalanced and high-dimensional datasets by modifying the ACO using multi-objective instead of the single-objective fitness function.

The BAT-based method is motivated by the echolocation behaviour of bats. *In the BAT-based method,* a composite variant of the BAT and an enhanced PSO algorithm is presented to improve the performance of the system [81]. The inclusion of the PSO algorithm was to reinforce the convergence power of the hybrid algorithm. A binary BAT algorithm was leveraged for feature selection for steganalysing of images [82]. An enhanced BAT (EBat) algorithm was presented by [83] to address the challenge of local optima trapping based on a special mutation operator that enhances the diversity of the standard BAT method.

The GSA-based method is motivated by Newton's law of universal gravitation. *In the GSA-based method* [84], presented a strategy where a piecewise linear chaotic map is explored for feature selection. In [85], GSA algorithms were enhanced for improving the performance of the conventional gravitational search algorithm for optimal FS misclassification task.

The FA-based method is inspired by the optical association in between fireflies, where extraordinary outcomes is accomplished by the working action and cooperation of low-performance agents. *In the FA-based method,* [86] presented a return-cost-based binary FA-based FS, which yields a variety of techniques to forestall premature convergence and increase the accuracy of the model. In [18], prevention of trapping in local optimization and an enhanced convergence is achieved by modifying the standard FA. In [87], the FA-based FS method is employed for classifying Arabic texts based on an SVM classifier [88]. Put forward an FA-based strategy for detecting network intrusion by utilizing the composition of filter-based and wrapper-based FS techniques.

The COA-based method is inspired by the extraordinary way of life of the cuckoo species of bird attitude of laying eggs and reproducing. *In the COA-based method,* [89] utilized a COA-based FS technique to enhance the classification of cancer classification data by first eliminating the redundant features and then selecting the final features using integration wrapper-based FS and the COA algorithm. In [90], the COA was enhanced to aid the quick diagnosis of disease. [91] employed a composition of the COA and neural network during the feature selection task for the detection and classification of heart disease.

The SSA-based method is motivated by the swarming behaviour of salps during their movement and scavenging in the seas. *In the SSA-based method,* [92] presented a hybrid optimization method that integrates the salp swarm algorithm with PSO to enhance the efficacy of the exploration and the exploitation steps in FS. [93] put forward the SSA feature weighting method for the prediction of the presence of Parkinson, heart and liver disease. In [94], a composition of an enhanced SSA and a local search algorithm is presented to address sparsity and high dimensionality of data for the FS. Other notable works in the literature based on the SSA approach can be found in [95, 96].

The WOA-based method is motivated by the hunting characteristic of humpback whales. *In the WOA-based method* [97], a synthesized WOA alongside a simulated annealing algorithm is presented for FS to reinforce the exploration phase by finding the most promising regions. In [98], a tournament and roulette wheel selection strategy with hybrid and mutation operators are employed to upgrade the exploration and exploitation of the search process based on WOA. [99] put forward a frequency-based filter FS approach which eliminates irrelevant features based on the WOA algorithm.

The GWO-based method is inspired by the natural hunting method of a pack of grey wolves. Grey wolves have an extremely intriguing behaviour. They frequently live and move in a pack and follow an exceptionally inflexible social hierarchy of strength and dominance. At the top of the hierarchy are the leaders referred to as the alphas who dictate rules that the group must obey. Immediately after the alphas are the betas who ensures the alphas

orders are obeyed and are predestined to succeed the alpha. The subset of other wolves controlled by the leading wolves are referred to as omega. Deltas are the remaining wolves who neither belong to the category of alpha, beta or omega. *In the GWO-based method,* [100] introduced a binary model of the GWO which chooses the ideal feature subset for classification tasks. The work constrained the position of the wolves only to binary values by modelling it in a discrete space to choose between selecting or discarding a given feature in the dataset. A multi-strategy ensemble GWO was introduced for FS to upgrade the standard GWO-based technique in [101]. In [102], a mutation operator is proposed to mitigate the selection of redundant and irrelevant features based on the GWO technique.

An extensive treatment of the swarm-based feature selection method and its categories can be found in the published article by Rostami et al. [57].

4.1.2 Evolutionary-based algorithm (EBA)

The evolutionary-based algorithm is sometimes sub-categorized under the swarm-based algorithm as their nature of behaviour is similar. Also, most recent works usually combine algorithms from both the SI and EBA to achieve optimal performance during the classification task. Some examples are the genetic algorithm (GA), differential evolution (DE), amongst others.

GAs are advanced algorithms based on the mechanics of biological and natural genetics, and they are mostly utilized for generating high-quality solutions for search and optimization issues based on the intuition of biologically inspired operators. [103] put forward a hybrid approach to determine the most suitable feature subset combined with a versatile neuro-fuzzy inference system for forecasting future electrical energy interest. A modified variant of the GA called MGA alongside a deep neural network was put forward for forecasting patients' demand for different essential resources in the outpatient department in hospitals [104]. A novel GA model was presented by [105] for generating and recognizing children's activities based on environmental sound. GARS, a GA-based algorithm for identifying a robust subset (GARS) and applicable for multi-class and high-dimensional datasets, is presented by [106]. It yields a high classification accuracy with reasonable execution time while taking a computation.

DE, which was presented by Storn and Price [107], is a composition of a parallel direct search technique in which search is executed in large, complex and multi-modular scenes to yield optimal solutions for objective or fitness function (FF) of an optimization problem. The DE algorithm performs mutation, crossover and selection operations. DE was put forward to mainly address the major limitation of the GA, which to be specific is the absence of local search. Hence, their primary difference is in the genetic selection operators. [108] presented an upgraded multi-objective DE algorithm to enhance classification accuracy and eliminate noisy and redundant features. The same authors put forward a novel multi-objective DE to enhance the performance of the clustering algorithm [109]. [110] proposed a self-adaptive DE algorithm called SaDE to address intrusion detection problems in wireless sensor networks (WSN). In [70], a multi-objective feature selection method called binary differential evolution with self-learning (MOFS-BDE) based on the multi-objective feature selection approach is presented. An evolutionary computation-based technique which is a hybrid multi-objective FS was presented by [111] to identify and select a small subset of features and achieve higher prediction results compared to utilizing all features.

As noted in the earlier section, some researchers combined multiple methods from SI, EBA and others to get a better performing model, for instance the ensemble method, which combines several ML techniques into one predictive model to decrease variance (bagging), bias (boosting) or improve predictions (stacking). Hence, improve the accuracy by combining the output of many weak learning classifiers. In improving the accuracy problem, the authors in [112] proposed a novel approach of hybrid model (BBO-bagging) for feature selection and classification. They employed a hybrid combination of nature-inspired algorithms. That is, biogeography-based optimization (BBO), particle swarm optimization (PSO) and genetic algorithm (GA), as a feature selection technique with the ensemble classifier to achieve an optimal text classification. They trained and tested the extracted features on six classifiers, namely: K-nearest neighbour (kNN), random forest (RF), support vector machine (SVM), Naïve Bayes (NB), decision tree (DT) and ensemble (Bagging). Based on the obtained results, their analysis demonstrated that the performance of (BBO) as a feature selection technique is better than independently using the (GA), (PSO) and the (BBO). Belazzoug et al. [113] proposed a new wrapper improved sine cosine algorithm (ISCA) with a combination of the information gain (IG) filter to avoid early convergence and reduce the large dimensionality challenge. The efficiency of this method was validated by employing nine text collections consisting of popular benchmark datasets. Based on the performance measures, the experimental results showed the ISCA performed higher compared to the original SCA algorithm. The ISCA used a few parameters set that let the proposed algorithm to be quite flexible and straightforward to apply to a broad spectrum of search problems. Likewise, their proposed algorithm may be combined with other search algorithms to get better performance.

4.1.3 Trajectory-based algorithms (TBAS)

Trajectory classification assists in understanding the character of objects being monitored. However, the raw trajectories might not yield satisfactory classification results. Hence, features are extracted from raw trajectories to enhance classification results [114]. Also, all the extracted features may not be helpful for classification. Therefore, an automatic selection scheme is vital for finding optimal features from the pool of handcrafted features such as used by genetic algorithms and random forests (RF). Trajectory-reliant models are sometimes classified using random forest (RF)-based classifier and then compared with a support vector machine (SVM). Detecting abnormal trajectories is a critical task in research and industrial applications. Industrial applications in video surveillance, maritime, smart urban transportation and climate change domains have attracted significant attention in recent times [115]. The trajectory-based FS method is still gaining ground and more research needs to be done to understand how it processes data. A relative comparison of studies that have applied the metaheuristics-based feature selection approach for text classification is given in Table 3.

The number of studies and the percentage of publication per publication date is shown in Table 4

The linear distribution of publication forecast in the year under review is presented in Fig. 7.

Figure 7 depicts the distribution of the publications linear forecast between year 2015 and June 2021. The R-square (R2) explains the accuracy of linear forecast on the reviewed articles concerning FS which is 67.23%. It is clear from Fig. 7 that the published articles increase annually. That means the topic attract more researchers yearly. Therefore, many solutions were proposed to the issue of feature selection optimization methods for optimal text classification.

Additionally, Table 3 discusses recent work spanning the year 2015 to 2021. Other forms of algorithms have been classified in some cases under SI and in other cases as EBA. For instance, the ***Pigeon-Inspired Optimization*** (PIO) algorithm is an intelligent algorithm spurred by the behaviour of pigeons where every pigeon of the swarm has a position, a speed, and an individual best historical position, as per its movement in the search space. PIOs have reportedly performed well in solving continuous optimization problems [133]. A discrete pigeon-inspired optimization algorithm that employs the Metropolis acceptance criterion of simulated annealing algorithm was put forward by [133] to address large-scale travelling salesman problems. They improved the discrete PIO exploration ability by developing a new map and compass operator with a comprehensive learning ability. The algorithm reinforces its capability to escape from premature convergence by utilizing the Metropolis acceptance criterion to decide whether to accept newly produced solutions. Duan and Qiao in [134] presented a PIO which served as an intelligence optimizer for addressing air robot path planning problems. The algorithm improved the convergence speed and also enhanced the superiority of global search in diverse use-cases. A hybrid algorithm that is fast, stable, and able to universally optimize the maximum power point tracking algorithm was presented by [135]. The algorithm is a composition of a new pigeon population algorithm called parallel and compact pigeon-inspired optimization (PCPIO) with maximum power point tracking (MPPT), which can address the problem MPPT cannot reach the near-global maximum power point. The quadrotor swarm formation control problem was addressed by [136] using a binary pigeon-inspired optimization (BPIO) model. The model solves the combination problem in the binary solution space using a special fitness function to avoid a crash and converge quickly.

The ***Fish Migration Optimization (FMO)*** algorithm, inspired by migratory greying, incorporates migration models and swim into the optimization process [137]. The binary fish migration optimization is a variant of FMO with the capability of converging quickly. FMO guides the evolution of the fish swarm (similar to PSO) based on the global optimal solution by utilizing the parameter to help the FMO carefully search the known space. To address the challenge of stagnation and falling into local traps, [137] proposed an advanced binary FMO. The algorithm improved the search ability of the BFMO by using the transfer function to map the continuous search space to the binary space.

Other recent work by [138] addresses the knapsack problem by utilizing a binary gaining sharing knowledge-based optimization algorithm. The ***Gaining Sharing Knowledge-based (GSK)*** optimization algorithm addresses binary optimization problems based on the concept of acquisition and sharing of knowledge of humans during their lifetime. The list of algorithms is all-encompassing as diverse metaheuristic-based optimization algorithms are coined by researchers daily based on the behaviour of the concept they intend to use for their algorithm. Some of them are the ***Binary Monkey Algorithm*** [139], ***discrete shuffled frog leaping algorithm*** [140], amongst others.

4.2 Evaluation measures

Evaluation of a predictive model is a critical phase in the classification task. This is after the model has been built and trained on some data. The modeller's concern becomes finding out how well the model is doing, how useful is the model, are more features needed, is there a need to train the

Table 3 Comparison of related studies on metaheuristic-based feature selection method for text classification

Reference used for the study	Feature selection methods	Dataset	Classification algorithms	Performance and evaluation methods	Contribution	Shortcomings
A novel community detection-based genetic algorithm for feature selection [116]	First, the similarities of the feature are calculated. In the second step, the features are classified by community detection algorithms into clusters. Third, the features are picked by a genetic algorithm	Nine benchmark classification problems were analysed in terms of performance	It used a genetic algorithm based on community detection. The selected methods are based on PSO, ACO, and ABC algorithms	Comparing the performance of the proposed method with three new feature selection methods based on PSO, ACO, and ABC algorithms on three classifiers showed that the accuracy was on the average of 0.52% higher than the PSO, 1.20% higher than ACO, and 1.57 higher than the ABC algorithm	The proposed genetic algorithm approach takes cognizant of the correlation between the selected features, hence preventing the selection of redundant features and significantly improving the predictive model's performance	To optimize the selected parameters, there is a need to repeatedly set parameters, generate a number of predictions with distinct combinations of values and then evaluate the prediction accuracy to select the best parameter values. As a result, choosing the best values for the parameters is an optimization problem
Comparison on feature selection methods for text classification [117]	Typical feature selection methods for text classification, alongside a comparison of experiments on four benchmark datasets, was conducted to compare the effectiveness of twenty typical FS methods	Four datasets achieved from the UCI repository are utilized in the comparison experiments. The four datasets are named as CARR, COMD, IMDB and KDCN, respectively	It uses MOR and MC-OR for text classification. Likewise, it applied the unsupervised term variance (TV), term variance quality (TVQ), term frequency (TF) and document frequency (DF) for efficiency and high classification accuracies	It performance is of the typical feature selection methods	The result of this paper gives a guideline for selecting appropriate feature selection methods for text classification academic analysis or real-world text classification applications	MOR and MC-OR are both the best choices for TC. However, the formulas of the two methods are relatively complex
Novel approach with nature-inspired and ensemble techniques for optimal text classification [112]	Biogeography-based optimization (BBO) with ensemble classifiers, genetic algorithm (GA) and particle swarm optimization (PSO)	Ten text datasets from UCI repository {tr11, tr12, tr21, tr23, tr31, tr41, tr45, oh0, oh10, oh15}, Real-time dataset from MOA, including Scientific documents, News and Airlines dataset of 539,384 records	Naïve Bayes (NB), K-nearest neighbour (kNN), support vector machine (SVM), random forest (RF), decision tree (DT) and ensemble classifier	The average precision was 83.87 with 70.67 recall. The average accuracy was 85.16 with a 76.71 average F-measure	The proposed hybrid BBO algorithm selects optimal subset of features The algorithm was tested on real-time dataset of airlines. Thus, depicting its feasibility for solving real-world problems	The proposed approach used imbalanced data leading to irregularities in the accuracy and F-measure of some of the dataset during performance analysis
Automatic text classification using machine learning and optimization algorithms [77]	The approach is based on the artificial bee colony algorithm with a sequential forward selection algorithm (SFS), where the selection technique utilizes a modest greedy search algorithm	Reuters-21578, 20 Newsgroup and real dataset	Machine learning-based automatic text classification (MLearn-ATC) algorithm based on probabilistic neural networks (PNN)	A precision of 0.847, a recall of 0.839, an F-measure of 0.843 and an accuracy of 0.938 was obtained on Reuters. A precision of 0.896, a recall of 0.825, an F-measure of 0.859 and an accuracy of 0.937 was obtained on 20 Newsgroup. A precision of 0.897, a recall of 0.845, an F-measure of 0.870 and an accuracy of 0.961 was attained on real dataset	The proposed algorithm outperformed Naive Bayes (NB), K-nearest neighbour (KNN), support vector machine (SVM) and probabilistic neural network (PNN) when a comparative analysis that measured performance was carried out Authors claimed that the proposed algorithm utilizes the minimum time and memory while performing the task	The accuracy of the algorithm was verified on particle swarm optimization (PSO), ant colony optimization (ACO), artificial bee colony (ABC) and firefly algorithm (FA) only. Nevertheless, it performance cannot be generalized as it was not compared with other optimization methods

Table 3 (continued)

Reference used for the study	Feature selection methods	Dataset	Classification algorithms	Performance and evaluation methods	Contribution	Shortcomings
Optimized deep belief network and entropy-based hybrid bounding model for incremental text categorization [118]	Entropy-based FS infused with a feature extraction process using a vector space model (VSM) which extracts the TF-IDF and energy features	20 Newsgroups and Reuters dataset	Grasshopper crow optimization algorithm (GCOA) and deep belief network (DBN)	A precision of 0.959, 0.959 recall and an accuracy of 0.96 were reported	The proposed algorithm provides better performance for incremental text categorization when compared with existing algorithms	The proposed algorithm was not compared alongside other evolutionary algorithms. Hence, its performance may not give the same result when compared to other known systems
Optimal feature subset selection using hybrid binary Jaya optimization algorithm for text classification [119]	Composition of wrapper-based binary Jaya optimization algorithm (BJO) and filter-based normalized difference measure (NDM)	WebKB dataset, SMS dataset, BBC dataset and 10Newsgroup dataset	Multinomial Naïve Bayes (NB) and linear support vector machine (SVM)	No set values given. Graphs were used to depict the superiority of the proposed NDM-BJO when compared with existing NB and SVM classifiers for the four categories of datasets used	Proposed a new hybrid feature selection method called the normalized difference measure and binary Jaya optimization algorithm (NDM-BJO) to reduce the high-dimensional feature space of text classification problem	The evaluation metrics was based on accuracy and F_1^{Macro}. However, much uncertainty still exists if the proposed algorithm will outperform existing system when other metrics such as precision or recall is used to evaluate the efficacy of the system
Optimization of multi-class document classification with computational search policy [120]	Cuckoo optimization (CO), firefly optimization (FO), and bat optimization (BO) algorithms - correlation-based feature subset filter	News documents	J48 and support vector machine (SVM)	The accuracy for J48 for CO, FO and BO are 92.03%, 90.55% and 90.23%, respectively The accuracy for SVM for CO, FO and BO are 87.22%, 89.60% and 87.22%, respectively	Proposed model took the advantage of nature-inspired-based metaheuristic algorithms, which provided advanced nature search for nonlinear complex problems	More classifiers need to be set up with computational search policies and their effects measured
An improved sine cosine algorithm to select features for text categorization [113]	Improved sine cosine algorithm (ISCA)	Reuters-21578 (Re0), La1s, La2s, Oh0, Oh5, Oh10, Oh15, FBIS, tr41	Naïve Bayes (NB)	The average precision, recall and F-measure are 82.32, 82.89 and 82.22, respectively	Proposed ISCA algorithm which is statistically significant than Obl-SCA, weighted-SCA and ACO algorithms	Proposed ISCA is statistically weak for some other algorithms such as GA, LevySca, SCA and MFO. Thus, limiting its generalization for conclusion if it improves the performance of categorization task in a larger setting
Text feature space optimization using artificial bee colony [73]	Artificial bee colony (ABC)	Reuters-21578	support vector machine (SVM) Naïve Bayes (NB) and k-nearest neighbours (KNN)	The average accuracy, precision, recall and F-measure on SVM are 95.07%, 84.75, 83.74 and 96.08, respectively On NB are 92.23%, 83.04, 81.96 and 82.48, respectively On KNN are 87.37%, 78.91, 77.25 and 78.04, respectively	Proposed the ABC, a metaheuristic-based algorithm for improved performance in text classification	Complexity in determining the control parameters or hyperparameters for the algorithm

Table 3 (continued)

Reference used for the study	Feature selection methods	Dataset	Classification algorithms	Performance and evaluation methods	Contribution	Shortcomings
New hybrid method for feature selection and classification using metaheuristic algorithm in credit risk assessment [121]	An exploration of new advanced hybrid feature selection has been proposed to deal with these problems	The new proposed algorithm utilizes the dataset from unique client identifier (UCI) repository of machine learning credit to estimate the performance	A metaheuristic of imperialist competitive algorithm with modified fuzzy min–max classifier (ICA-MFMCN)	Statistical test results show that the available data support the hypothesis of searching for reliability level of 1%	Fast algorithms performance to future ranking and also optimization capabilities of an ICA	Lack of rapid filtering technique to reduce the search space
Artificial bee colony algorithm for feature selection and improved support vector machine for text classification [74]	Based on artificial bee colony feature selection (ABCFS) algorithm	Reuters-21578, 20Newsgroup corpus and Real datasets	Support vector machine (SVM) and improved SVM (ISVM)	The average precision, recall, F-measure and accuracy on Reuters are 0.675, 0.702, 0.679 and 0.829, respectively On 20 Newsgroup are 0.701, 0.723, 0.710 and 0.822, respectively On real dataset are 0.840, 0.797, 0.817 and 0.835, respectively	Proposed ABCFS which enhances the accuracy of text document classification	Proposed algorithm requires a high computational time and complexity Verified only on SVM and an improved. Hence, it is unclear if the performance of the algorithm can be generalized on other state-of-the-art classifiers
A modified multi-objective heuristic for effective feature selection in text classification [122]	Modified artificial fish swarm algorithm (MAFSA)	OHSUMED	Support vector machine (SVM), AdaBoost classifiers and Naïve Bayes	Average precision of MAFSA is 2.27% better than artificial fish swarm algorithm (AFSA)	Proposed MAFSA which is an improvement over AFSA for feature selection and better text classification	Performance metrics not descriptive enough
An ACO–ANN-based feature selection algorithm for big data [76]	Ant colony optimization (ACO)	Reuters-21578	Artificial neural network (ANN)	The average precision, recall, macro F-measure, micro F-measure and accuracy are 77.34, 80.14, 79.01, 89.87 and 81.35, respectively	Proposed ACO algorithm is a subset of the hybrid algorithm which has the capability to congregate promptly since it has effective search ability in the problem state space, thus allowing the efficient determination of minimal feature subset	The performance of the proposed algorithm cannot be generalized as verification was not done on standard classifiers
Competitive particle swarm optimization for multi-category text feature selection [123]	Continuous particle swarm optimization (PSO) algorithm	RCV1 and Yahoo collections	Multi-label Naive Bayes (MLNB) and extreme learning machine for multi-label (ML-ELM)	One-error for MLNB EGA + CDM, bALO-QR and CSO are 3.75, 2.31 and 2.94m respectively Multi-label accuracy for MLNB are 3.19, 2.75 and 3.06m respectively	Proposed a process for estimating the relative effectiveness of the PSO based on the fitness-based tournament of the feature subset in each iteration hybridized approach addresses degenerated final feature subsets	The performance of the proposed algorithm cannot be generalized as verification was not done on standard classifiers The proposed PSO was designed for multi-label text feature selection. It was not tested on single-labelled text

Table 3 (continued)

Reference used for the study	Feature selection methods	Dataset	Classification algorithms	Performance and evaluation methods	Contribution	Shortcomings
A new approach for text documents classification with invasive weed optimization and Naive Bayes classifier [124]	Invasive weed optimization (IWO) and Naive Bayes (NB) classifier (IWO-NB)	Reuters-21578, WebKb, and Cade 12	Naive Bayes (NB)	The precision, recall, F-measure, AUC, accuracy and error rate on Reuters are 0.6632, 0.6925, 0.6775, 0.6894, 0.7012 and 0.2988, respectively On WebKb are 0.6548, 0.7136, 0.6829, 0.6914, 0.7265 and 0.2735, respectively On Cade 12 are 0.6984, 0.7214, 0.7097, 0.7058, 0.7045, 0.2955	Proposed a hybrid of IWO algorithm and NB classifier for improving the performance of document classification	The performance of the proposed algorithm cannot be generalized as verification was not done on all standard classifiers
Particle swarm optimization-based two-stage feature selection in text mining [125]	Correlation (CO), information gain (IG), gain ratio (GR), symmetrical uncertainty (SU) and particle swarm optimization (PSO)	Reuters-21578 R8 dataset	Naïve Bayes (NB)	Average accuracy on CO is 88.74%, on IG is 89.52%, on GR is 87.83% and on SU is 89.34%	Proposed algorithm eliminates useless features and reduces the search space for enhanced performance during the categorization task	Requires increased computational resources and complexity Requires more numbers of features Approach requires further work such as using a different fitness function or a multi-objective search approach
A text feature selection method based on the small world algorithm [126]	Information gain (IG) and Chi-square statistics (CHI) Optimization of candidate features: Small world algorithm (SWA)	Reuters 21,578 Classic Corpus, Chinese Fudan Corpus	K-nearest neighbours (KNN) and support vector machine (SVM)	Aggregated accuracy on Reuters improved by an average of 2.3% when using the IG or CHI and SWA optimization Aggregated accuracy on Fudan improved by an average of 5.3% when using the IG or CHI and SWA optimization	The proposed algorithm minimized the dimension of feature vector and the complexity ultimately increasing the accuracy rate A local short-range search algorithm was used to improve performance of text classification	There was no optimization on the parameter setting of the SWA thus making some portion of the results to be inconclusive The proposed SWA algorithm has no optimal number of iterations due to the lack of a mechanism for parameter settings
An improved flower pollination algorithm with AdaBoost algorithm for feature selection in text documents classification [127]	Flower pollination algorithm (FPA)	Reuters-21578, WebKb, and Cade 12	AdaBoost	The precision, recall, F-measure and accuracy on Reuters are 77.94, 69.32, 72.77 and 70.35, respectively On WebKb are 76.54, 69.94, 71.95 and 69.48, respectively On Cade 12 are 76.94, 71.24, 73.81 and 69.89, respectively	Proposed model shows a significant reduction in the size of features as well as the similarity between the categories of weight and the distance between the words when compared with other models	Proposed model is dependent on the parameter values making it less efficient when choosing the feature weights

Table 3 (continued)

Reference used for the study	Feature selection methods	Dataset	Classification algorithms	Performance and evaluation methods	Contribution	Shortcomings
An improved k-nearest neighbour with crow search algorithm for feature selection in text documents classification [128]	Crow search algorithm (CSA)	Reuters-21578, WebKb and Cade 12	K-nearest neighbour (KNN)	The precision, recall, F-measure and accuracy for KNN on Reuters are 76.34, 69.47, 72.74 and 68.32, respectively On WebKb are 77.35, 68.24, 72.51 and 70.64, respectively On Cade 12 are 75.48, 69.58, 72.41 and 72.23, respectively	Proposed model is more accurate in classification than the standard KNN with a greater F-measure Proposed model gave a higher accuracy of 27% when compared to KNN	Proposed model have the drawback of optimal feature selection during the classification task
Multi-label text classification using optimized feature sets [129]	Wrapper-based hybrid artificial bee colony and bacterial foraging optimisation (HABBFO)	Reuters dataset	Artificial neural network (ANN)	The precision, recall and hamming loss for KNN are 89.85, 88.89 and 35.45, respectively For ANN are 94.82, 93.79 and 20.45, respectively	The proposed multi-label classifier performs better than standard KNN algorithm when evaluated in terms of precision, recall and hamming loss	Proposed feature selection model was verified on KNN and ANN classifier only. It generalization on other classifiers using the authors proposed algorithm is undefined
Feature selection for text classification using genetic algorithms [130]	Genetic algorithm (GA)	20Newsgroups, Reuters-21578	Naive Bayes (NB), nearest neighbours (KNN) and support vector machines (SVMs)	F-measure on Reuters for KNN, SVM and NB are 0.931, 0.946 and 0.863, respectively F-measure on 20Newsgroup for KNN, SVM and NB are 0.931, 0.879 and 0.946, respectively	Proposed algorithm allows search of a feature subset such that the performance of classifier is best It allows finding a feature subset with the smallest dimensionality which yield higher accuracy in classification	Algorithm needs to be verified on evolutionary and metaheuristic algorithms or hybrid solution to improve textual document classification
Metaheuristic algorithms for feature selection in sentiment analysis [131]	The study compares feature selection in text classification based on traditional and sentiment analysis methods	The proposed dimension reduction strategy was to reduce the size of large capacity training dataset	It applied metaheuristic method such as genetic algorithm, particle swarm optimization (PSO) and rough set theory	The result of the research in traditional text classification found that ACO was able to obtain optimum feature subset compared to GA	The result shows metaheuristic-based algorithms have the potential to be perform in sentiment analysis	The main challenges in the sentiment classification are overlapping of features, large size dimension, and irrelevant elimination
A new and fast rival genetic algorithm for feature selection [132]	The study put forward a new rival genetic algorithm (RGA) to improve the performance of GA for feature selection	The study used twenty-three (23) benchmarked dataset encompassing UCI machine learning repository dataset and Arizona State University dataset	Not stated	Average accuracy was 0.9579 Average FSR is 0.4386	A competition strategy and dynamic mutation rate was used to enhance the performance of the GA A fast RGA was presented to enhance the computational effort of RGA	Future work require testing the efficiency of the RGA on other unexplored classification tasks such as electromyography signals, detection and diagnosis of strokes and other diseases

Table 4 Distribution of publications per year

Year	Number of studies	Publications in percentage
2015	15	7%
2016	20	10%
2017	25	12%
2018	30	15%
2019	35	18%
2020	45	23%
2021	30	15%
Total	**200**	**100%**

model to improve its overall performance, can its performance be generalized, etc.

In the general classification task, the overall outcome is usually measured using the following:

- True positives mean that the model's prediction is positive and in reality, it is positive.
- True negatives mean that the model's prediction is negative and in reality, it is negative.
- False positives mean that the model's prediction is positive and in reality, it is negative.
- False negatives mean that the model's prediction is negative and in reality, it is positive.

A confusion matrix is often used to plot and display the outcome in a matrix format. The outcomes postulate the metrics used for evaluation. Some of the metrics often used are precision, recall, accuracy, specificity, F-measure, mean squared error, area under curve, logarithmic loss, ROC (Receiver Operating Characteristics) curve, mean absolute error, etc. The metric to use for evaluation depends hugely on the task at hand.

The ***Precision*** metric postulates the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

The ***Recall*** metric postulates the percentage of positive instances out of the *total actual positive* instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

The ***Accuracy*** metric postulates the ratio of the number of predictions that are correct to the total number of input samples.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions performed}} \quad (3)$$

The ***Specificity*** metric postulates the percentage of negative instances out of the *total actual negative* instances.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (4)$$

The ***F-measure*** metric postulates the harmonic mean of precision and recall.

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

The ***Mean Squared Error*** essentially characterizes the average of squared differences between the actual output and the predicted output.

The ***Area Under Curve*** estimates the capability of a binary classifier to discriminate between positive and negative classes.

The ***Logarithmic Loss*** estimates the model's performance where the prediction input is a probability value in the range of 0 and 1.

The variants of evaluation metrics are quite exhaustive, and thus, only the main ones were briefly discussed. In summary, evaluation measures delineate the performance of a model. The intuition behind the development of predictive models works on a constructive feedback principle. A model is fabricated, followed by getting feedback from metrics, improvements are made and repeated until the desired outcome is accomplished. From the main papers reviewed in Table 3, most researchers focused on using

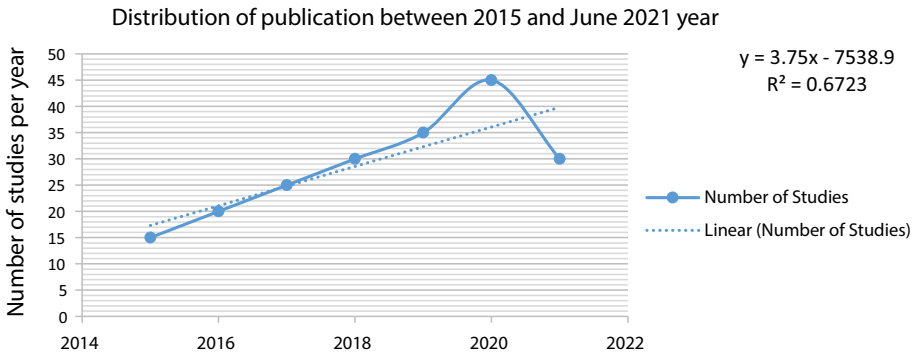

Fig. 7 Distribution of the Publications in the Years under Review

accuracy, precision and recall metrics. Several researchers estimated the F-measure too.

4.3 Datasets

Dataset is the core of every predictive model. Some of the frequently used models for testing and training textual data as found in the literature are 20 Newsgroups, Reuters-21578 and so forth. ***20 Newsgroups*** was developed by Carnegie Mellon University. It is a collection of about 18,000 newsgroups posts on 20 different topics. It has several sub-categories of datasets each of which is split into two sections; one is used for training and the other is utilized for testing the model [141]. The rule for splitting the training and test subsets usually depends on the posting date been a previous or following a particular day. The 20 Newsgroups dataset has become popular for researching predictive models. The news is categorized based on its contents. The ***Reuters-21578*** is a text dataset and its documents are organized in a hierarchical structure containing 21,578 news articles, each belonging to a category or more through different points of view [142]. Based on the analysed data, most of the works in the literature used the Reuters-21578, 20 Newsgroups amongst others.

Dataset-related challenges are the primary reason why optimal performing real-model seems unachievable. To avoid overfitting the model to the data, small datasets require models that have high bias or low complexity.

4.4 Research gaps

4.4.1 Confusion of evaluation metrics to be used

The assessment of the performance of a model discloses how well it executes on unseen data. In practice, making predictions on future data is what the model is built for. Thus, it is a major problem that the predictive model wants to solve. Therefore, there is a dire need to understand the context prior to choosing a metric in light of the fact that each model attempts to address a problem with a different objective using a different dataset. For instance, most researches focused on precision, recall and accuracy. However, precision and recall are effective metrics mostly in cases where classes are not evenly distributed.

4.4.2 Specificity of data types and domain agreement

Data types narrowed to specific domains to solve generalised problems in the domain in view are a gap in the literature. Having a grouping of specific data types to work in an agreed and generalized domain will lead to a fitting and good predictive model with little or no challenge when deployed on unseen data. Such lapses in the field of data mining have continued to produce models that only memorize and fail to generalize accurately on unseen data. Such an idea was presented by [143] in the bioinformatics domain. However, other domains remain unexplored in such regard.

4.4.3 Dataset issue

In most of the reviewed articles, it is observed that a considerable number of researchers used the Reuters-21578, meanwhile others used different datasets. Subsequently, the performance evaluation is subject to the specific dataset, classes and classifiers used, which brings about the challenge of a benchmark. A comparison between the feature selection algorithms needs to be carried out using a single dataset and the same group of classifiers. In that way, a standard can be reached in terms of comparison of the distinct datasets using dissimilar metaheuristic algorithms.

4.4.4 Established benchmarks

There is an urgent need to institute an established benchmark. This is important so that the correctness of any model can also be validated through the use of benchmarks. The specific algorithm can then be quickly evaluated. The current way of selecting different datasets, classifiers and evaluation criteria by an individual researcher makes it almost impossible to ascertain which metaheuristic algorithm performs better than the others when classifying text even in the same domains.

4.4.5 Hybrid search issues

Previous works by Ghareb et al. [144]; Lee et al. [145] have established the notion of improving the search space through hybridizing with a filter using evolutionary-based algorithms. Likewise, there is still the issue of the fitness of a feature subset requiring improvement after modification. It results to wastage when performing computations fitness of the algorithm and evaluation. Although, the study by Lee et al. [123] attempted to bridge the gap by selectively applying a single operator to minimize the number of feature subsets to increase the number of times the fitness is improved. However, their approach was applied only to multi-label text feature selection. This is a prospective topic for research that will be recommended in the future work section.

4.4.6 Relevancy of a feature issue

In recent times, the relevancy of a feature was raised by some researchers [22], like how to measure the relevancy

of a feature to the data or the output. Many publications have presented various definitions and measurements for the relevance of a variable in feature selection [22, 35, 146]. There is a need for research exploration into resolving the issue of relevancy and irrelevancy of a variable in the feature selection process.

5 Lessons learnt during this review

i. As shown in Table 1 and the limitations highlighted in Table 2, feature selection is highly context and data-reliance. Therefore, there is no one-fits-all solution or one-stop for feature selection during the classification process. The strategy is to understand the process of each technique and deplore it when required.
ii. Each metaheuristic paradigm has its own set of merits and shortfall that makes it more suitable to a specific application. Nevertheless, finding the best-suited metaheuristic algorithm is a complex task as metaheuristic algorithms do not totally guarantee optimum solutions, which is due to the issue of theoretically establishing the efficiency of algorithms. Typically, studies rely on the empirical results to prove the same kind of solutions. Additionally, the task of designing some metaheuristic framework before its application for solving the problem in view may be so challenging.
iii. An important observation in the application of the metaheuristic-based approach is the discrepancy in finding remarkable solution to the problem at hand. Researchers validated their algorithms using different evaluation metrics and different datasets. This variation in the researcher's report makes it quite challenging to generalize the performance of one algorithm and the other. Successful deployment of standardized systems of a metric format will be helpful to a newcomer in the field. Especially to quickly look through algorithms that least perform and use the shortcoming to make progress faster in discovering new solutions that yield better results.
iv. As noted by Jiao and DU [143], a proficient approach to interpreting the performance values during comparison is to perform a thorough and rigorous analysis utilizing an identical testing dataset, identical training dataset and identical evaluation protocols, though they highlighted that such requirements might be practically difficult to satisfy. Notwithstanding, analysts should note that better performance measures may not guarantee better performance in practical, real-world applications as long as the comparison is not performed in the approach as mentioned earlier.
v. The FS ought to be envisioned as part of the training procedures. In the event that the FS procedure utilizes the entire dataset, and cross-validation is performed after that on the same entire dataset with selected features, the predictive performance has a high likelihood of being overestimated, as such procedures occur in a rigorous mathematical way. Subsequently, making analysts think that some information of the testing sample has slipped into the training dataset by helping to decide which features are selected. Therefore, it is safer to leave the testing sample out before the FS cycle during the evaluation stage [143].

6 Other issues and possible solutions

Other challenges in using metaheuristics for text classification are highlighted as follows.

i. **Time delay Processing** is still a weakness in the application of metaheuristic-based algorithms. For example, in ACO, due to dimension and significant data size problem, it often takes a long time to process [147]. Designing algorithms in a short amount of time is required to curb diverse classification problems used in real-world settings. Hence, research is urgently needed to construct process algorithms that can be fast in handling feature selection and classification.
ii. **Overlapping of features**: The increase in feature size often causes overlapping of features [131, 148]. It is necessary to consider potential metaheuristic techniques or hyper-heuristics optimization techniques that can minimize feature size by eliminating overlapping features during the development of the system.
iii. **High accuracy problem:** Many classification issues, such as real-world classifications, encounter low accuracy performance. While most researcher's claim that their algorithms achieves higher accuracy and outperform other existing algorithms, the lack of standard domain-specific metrics, a benchmark of evaluation and a dataset makes it difficult to conclude on their claims and findings. Also, in cases of practical use and significance of many classification problems, for example, in the real-time application of crime detection. Several algorithms need to be developed to tackle the challenges in the classification with high accuracy to really curtail the problem at hand, specifically, where the application

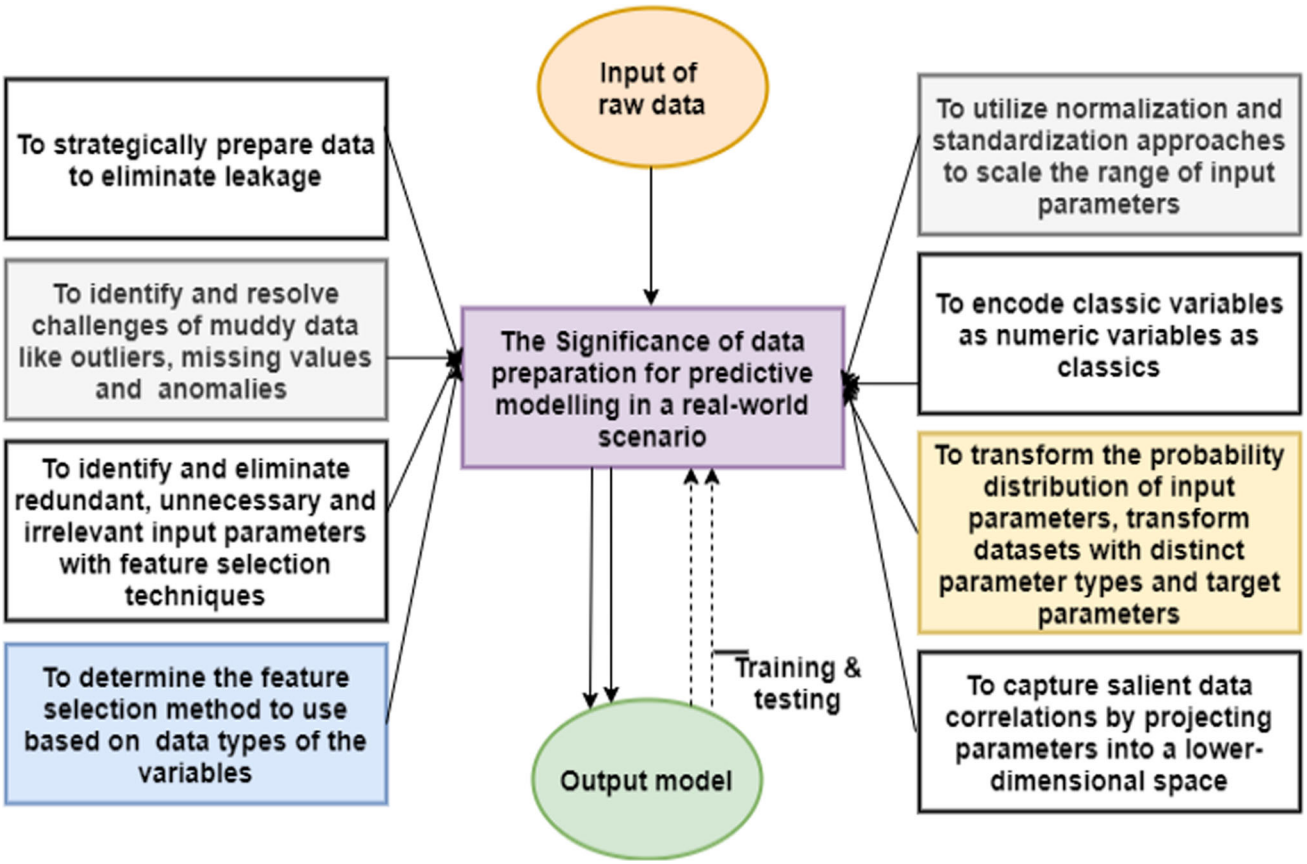

Fig. 8 Referential guide for developing good models

is in real-time. Hence, there is a need to address the challenge of applying metaheuristic-based algorithms that produce low accuracy during their application or performance.

iv. **Evaluation Challenges:** Taking cognizance of how a model generalizes on unseen data is a critical phase that must be overviewed in every predictive model's pipeline. In this manner, it is imperative to outwardly alongside visually inspect the data and the corresponding predictive model while assessing its performance. It is advisable to reduce the dataset into two dimensions and afterwards, plot the observations and decision boundary.

v. **Specificity and generalization of model challenge**: Each project is different because the specific data used in each undertaking are unique. Notwithstanding, the path to a good outcome is generally the same from project to project. This is alluded chiefly to the applied ML process [149, 150]. Knowing what data preparation techniques to use are often a difficult task. However, looking at the data preparation step regarding the context of the whole project makes it straightforward and more manageable.

Given the diverse methods, each of which may have their own setup and prerequisites. Nevertheless, the ML process steps before and after data preparation can help to inform what techniques to consider. Discovering how to best uncover the learning algorithms to the unknown underlying structure of the prediction problem in view requires a detailed data preparation process. This becomes less cumbersome and viable when the designer knows the types of data preparation to consider, the algorithms for each class of technique, how and when to configure techniques. A modelling manual/referential guide to developing a good model as a feature selection process is given in Fig. 8.

Thus, Fig. 8 shows the process of feature selection by highlighting the modelling manual/referential guide to developing a good model.

7 Challenges and future directions

i. No doubt, the current literature on metaheuristic-based feature selection is evolving. Therefore, future work can consider reducing some of the drawbacks of stand-alone methods by hybridizing

metaheuristics. While such methods may be demanding during the development of the framework, it will undoubtedly present very effective and satisfactory results. Such models will be more suited to problems with multiple local minima than exact methods that have higher chances of being stuck at the local optimum.

ii. Potential researchers who want to avoid the challenge of (i) may consider exploring the idea of hyper-heuristics-based feature selection optimization method. The hyper-heuristics-based optimization method can help decide on an optimum sequence of metaheuristic algorithms. Therefore, combining the advantages of each in obtaining the best feasible solution.

iii. The trajectory metaheuristics deal with only one solution at a time. In such a method, the search process explains a trajectory in the search space [18, 46–151]. In the last few years, there have been an emergency of many algorithms that do not entirely follow the paradigm of a pure metaheuristic. They combine algorithmic components originating from various optimization algorithms to provide outstanding solutions to real-world challenges. There are very few applications of the trajectory-based metaheuristics of optimizing feature selection for text classification. This is another area that future researchers may want to explore.

iv. As discussed in Section V and item (iii), there is a need to strike a balance and has a standard benchmark dataset, metrics for each category of algorithms which is another avenue for future work.

v. Future work may also consider expanding the study by Lee et al., [145] to enhance hybrid search and apply it to a single-labelled text using evolutionary-based algorithms. This is due to the outstanding outcome of it increasing the number of times the fitness is improved and without additional computations when used on multi-labelled text.

vi. A combination of deep learning, along with metaheuristic or hyper-heuristic methods, might produce the outstanding results, which is a potential area for future research.

vii. One of the major problems of the ML intrusion detection system (IDS) is the expensive computation time due to the incomplete, and unrelated features and redundant contents in the IDS datasets. To overcome such problems and ensure building efficient and more accurate IDS models, many researchers utilize pre-processing techniques like feature selection, normalization and a hybrid modelling technique is usually applied. Therefore, there is need to propose more future work on hybrid IDS modelling method with an algorithm for creating the IDS and feature selection with high predictive ability.

viii. One of the key issues observed in all the models developed in the literature to address network intrusion detection system (IDS) is the high number of false alert rate [152, 153]. Aside a high detection rate, a good IDS model should possess a very low false alert rate, and hence, more future work can be performed focusing specifically on reducing high false alert rate in metaheuristic models of feature selection.

ix. To maximize predictive ability, there is need to focus on metaheuristic optimization algorithm that will address many problems in modelling on feature selection and text classification. For instance, using Kernel partial least squares regression (KPLS) technique could optimize predictive ability [154].

x. Moreover, prospective area of research is to determine the control parameters or hyperparameters for metaheuristic algorithms. There are no enough works in the literature which explored that specific area. Hyperparameters for metaheuristic algorithms are an area that can help in testing different values of control parameters during the evaluation phase of estimating the viability of the algorithm [155]. The accuracy of network for a specific task greatly depends on the hyperparameters' configuration [156].

xi. Furthermore, hyper-heuristics is another green area of research that can help in resolving complex computational search and feature selection problems [157, 158]. The definition of hyper-heuristics was recently extended to mean a learning mechanism or search method for generating or selecting heuristics to address computational search challenges. Thus, more future work can be carried out to tackle complex computational search problems using hyper-heuristics models of feature selection.

8 Conclusion

Recent developments in knowledge discovery in information technology have put data mining as an extremely active and evolving area. Data mining helps human in finding, deciphering and interpreting hidden information from enormous raw data. Such research has brought about

text classification techniques that are vastly used for facilitating multimedia data processing in many applications, such as image tagging, e-mail processing, multimedia recommendation, and so forth. In addition, the surge of the amount of digital data from diverse sources such as web pages, social media, emails, online advertisements, blogs and e-libraries shows the improving value of text classification.

To an increasing extent, this rapid creation, share and exchange of data make undertaking the task of data analysis, extraction and knowledge retrieval very challenging. To be able to extract knowledge and gain insight from data, there is a need to first decrease the dimensionality of the data. Feature selection process is an indispensable data preparation phase that helps to reduce the dimensionality of data of a predictive model. However, it is a very complex and computationally demanding task which, if not appropriately performed defeats the main aim of extracting knowledge and the usability of any predictive model in real-world applications.

Feature selection is a significant task that enables the model to perform faster, eliminate noisy, less informative data, improve the model's precision and accuracy, remove redundant features, reduce overfitting of the model, and increase generalization on testing data. While the conventional feature selection techniques have been leveraged for classification tasks in the past few decades, they fail to optimally reduce the high dimensionality of the feature space of texts, thus breeding inaccurate and inefficient predictive models. Emerging technologies such as metaheuristic and hyper-heuristic optimization methods provide a new paradigm for feature selection because they produce impressive results which are accurate for optimal classification compared to conventional techniques. Metaheuristic methods can efficiently enhance the accuracy of computation demands, classification and storage; thus, it has been applied increasingly in diverse fields. However, little details are known on best practices for case-to-case usage of emerging feature selection methods. The literature continues to be engulfed with clear and unclear findings in leveraging the most effective method, which, if not performed accurately, alters precision, real-world-use feasibility and the predictive model's overall performance.

In this study, a systematic review of the metaheuristic-based feature selection methods for enhancing text classification was performed. The review answered many questions, such as the sub-field of metaheuristics, how it affects the accuracy of text classification, datasets, amongst others. Therefore, this paper provides a high-level snapshot of the research landscape in selecting metaheuristics, focusing on current progress made in the field and new areas to address for better solutions to feature selection challenges. This study is a matter of urgency due to the absence of precise details and subtleties on metaheuristic-based feature selection methods, which influences the accuracy, practicality and overall performance of predictive models. Hence, perceiving the impact, recognizing the effect and significance of FS in text classification, identifying the best techniques for selecting informative and relevant features from the context using metaheuristics methods implies researching, investigating and exploring the current literature to comprehend where each method stands at present.

Competitive performances on previous and current studies on the metaheuristics-based feature selection method were investigated. The review was then extended to additional related issues such as research gaps, lessons learned, as well as other issues and how they can be surmounted for the design of robust metaheuristic algorithms.

While proposing that metaheuristic methods can be employed in selecting features for text classification, one can also recommend using hybrid metaheuristics. More also, one can harness hyper-heuristics to provide an efficient strategy of dealing with highly complex optimizations challenges for feature selection in industrial and scientific-based domains for text classification.

Furthermore, the review indicates that using a method like metaheuristics-based optimization for feature selection and its hybridized version is a promising and fast-developing field. It can offer exciting opportunities and present many challenges. In conclusion, feature selection is an essential stage in text classification that should be studied comprehensively to navigate businesses towards a future with high performing algorithms to address real-world challenges.

Declarations

Conflict of interest All authors declare that there are no conflicting interests of whatsoever.

References

1. Malik PK, Sharma R, Singh R, Gehlot A, Satapathy SC, Alnumay WS, Nayak J (2020) Industrial Internet of Things and its applications in industry 4.0: State of the art. Computer Communications
2. Verma L, Lee SS (2011) Proliferation of wi-fi: Opportunities in ce ecosystem. In 2011 IEEE Consumer Communications and Networking Conference (CCNC) (pp. 213–217). IEEE
3. Zaidi S, Atiquzzaman M, Calafate CT (2020) Internet of Flying Things (IoFT): A survey. *Computer Communications*
4. Abiodun OI, Abiodun EO, Alawida M, Alkhawaldeh RS, Arshad H (2021) A review on the security of the internet of things: challenges and solutions. Wireless Personal Communications, 1–35
5. Zeimpekis D, Gallopoulos E (2006) TMG: A MATLAB toolbox for generating term-document matrices from text collections. In:

Kogan J, Nicholas C, Teboulle M (eds) Grouping multidimensional data. Springer, Berlin, Heidelberg, pp 187–210
6. Subramaniam A (2020) "What is Big Data? — A Beginner's Guide to the World of Big Data Awareness", Available: https://www.edureka.co/blog/what-is-big-data/. [Accessed: 24- Sept- 2020]
7. Hutchinson A (2019) Facebook Messenger by the numbers 2019, https://www.socialmediatoday.com/news/facebook-messenger-by-the-numbers-2019-infographic/553809/. [Accessed: 4- Jan- 2021]
8. Vega M (2020) "15+ Incredible Facebook Messenger Statistics in 2020", Available: https://review42.com/facebook-messenger-statistics/. [Accessed: 24- Sept- 2020]
9. Stancheva T (2020) "Crucial Twitter Statistics, Facts and Prediction in 2020", Available: https://review42.com/twitter statistics/#:~:text=Twitter's%20monthly%20active%20users%20amount,users%20are%20%E2%80%9Caffluent%20Millenials.%E2%80%9D/. [Accessed: 24- Sept- 2020]
10. Wikipedia (2021), Wikipedia Statistics. Available: https://en.wikipedia.org/wiki/Wikipedia:Statistics#:~:text=This%20is%20an%20information%20page.&text=Currently%2C%20the%20English%20Wikipedia%20includes,be%20analysed%20in%20many%20ways. [Accessed: 2- March- 2021]
11. Gogna A, Tayal A (2013) Metaheuristics: review and application. J Exp Theor Artif Intell 25(4):503–526
12. Dorigo M (1992). Optimization, learning and natural algorithms. PhD Thesis, Politecnico di Milano
13. Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning
14. Moscato P (1989) On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms. Caltech concurrent computation program, C3P Report, 826
15. Kennedy J, Eberhart R (1995) Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE
16. Price K, Storn RM, Lampinen JA (2006) Differential evolution: a practical approach to global optimization. Springer Science and Business Media
17. Alomari OA, Makhadmeh SN, Al-Betar MA, Alyasseri ZAA, Doush IA, Abasi AK, Zitar RA (2021) Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators. Knowl-Based Syst 223:107034
18. Zhang L, Mistry K, Lim CP, Neoh SC (2018) Feature selection using firefly optimization for classification and regression models. Decis Support Syst 106:64–85
19. Awadallah MA, Al-Betar MA, Hammouri AI, Alomari OA (2020) Binary JAYA algorithm with adaptive mutation for feature selection. Arab J Sci Eng 45(12):10875–10890
20. Hammouri AI, Mafarja M, Al-Betar MA, Awadallah MA, Abu-Doush I (2020) An improved dragonfly algorithm for feature selection. Knowl-Based Syst 203:106131
21. Mafarja M, Aljarah I, Heidari AA, Faris H, Fournier-Viger P, Li X, Mirjalili S (2018) Binary dragonfly optimization for feature selection using time-varying transfer functions. Knowl-Based Syst 161:185–204
22. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
23. Pereira RB, Plastino A, Zadrozny B, Merschmann LH (2018) Categorizing feature selection methods for multi-label classification. Artif Intell Rev 49(1):57–78
24. El-Kenawy ESM, Eid MM, Saber M, Ibrahim A (2020) MbGWO-SFS: modified binary grey wolf optimizer based on stochastic fractal search for feature selection. IEEE Access 8:107635–107649
25. Mansour NA, Saleh AI, Badawy M, Ali HA (2021). Accurate detection of Covid-19 patients based on feature correlated naïve bayes (FCNB) classification strategy. J Ambient Intel Hum Comput 1–33
26. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Cao B (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet 395(10223):497–506
27. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
28. Hall MA (2000). Correlation-based feature selection of discrete and numeric class machine learning
29. Welch BL (1947) The generalization of student's' problem when several different population variances are involved. Biometrika 34(1/2):28–35
30. Zhang Y, Dong Z, Phillips P, Wang S, Ji G, Yang J, Yuan TF (2015) Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Front Comput Neurosci 9:66
31. Gu Q, Li Z, Han J (2012) Generalized fisher score for feature selection. arXiv preprint
32. Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93
33. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, Nowe A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinf 9(4):1106–1119
34. Phuong TM, Lin Z, Altman RB (2006) Choosing SNPs using feature selection. J Bioinform Comput Biol 4(02):241–257
35. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
36. Spearman C (1987) The proof and measurement of association between two things. Am J Psychol 100(3/4):441–471
37. Saeys Y, Abeel T, Van de Peer Y (2008) Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 313–325). Springer, Berlin, Heidelberg
38. Blachnik M (2009) Comparison of various feature selection methods in application to prototype best rules. In: Kurzynski M, Wozniak M (eds) Computer Recognition Systems 3. Springer, Berlin, Heidelberg, pp 257–264
39. Bermejo P, Gámez JA, Puerta JM (2011) A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. Pattern Recogn Lett 32(5):701–711
40. Visalakshi S, Radha V (2017) A hybrid filter and wrapper feature-selection approach for detecting contamination in drinking water management system. J Eng Sci Technol 12(7):1819–1832
41. Wah YB, Ibrahim N, Hamid HA, Abdul-Rahman S, Fong S (2018) Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika J Sci Technol, 26(1)
42. Blum C, Roli A (2003) Metaheuristics in combinatorial optimization: overview and conceptual comparison. ACM Comput Surv (CSUR) 35(3):268–308
43. Sörensen K, Glover F (2013) Metaheuristics. Encycl Oper Res Manag Sci 62:960–970
44. Abd-Alsabour N, Ramakrishnan S (2016) Hybrid metaheuristics for classification problems. Pattern Recognit Anal Appl 10:65253
45. Blum C, Roli A (2008) Hybrid metaheuristics: an introduction. In: Blum C, Aguilera MJB, Roli A, Sampels M (eds) Hybrid metaheuristics. Springer, Berlin, Heidelberg, pp 1–30
46. Raidl GR (2015) Decomposition based hybrid metaheuristics. Eur J Oper Res 244(1):66–76

47. Blum C, Puchinger J, Raidl GR, Roli A (2011) Hybrid metaheuristics in combinatorial optimization: a survey. Appl Soft Comput 11(6):4135–4151
48. Blum C (2005) Ant colony optimization: introduction and recent trends. Phys Life Rev 2(4):353–373
49. Dorigo M, Stutzle T (2004) Ant colony optimization. The MIT Press, Cambridge, MA
50. Al-Betar MA, Awadallah MA, Abu Doush I, Alsukhni E, Alkhraisat H (2018) A non-convex economic dispatch problem with valve loading efect using a new modifed β-hill climbing local search algorithm. Arab J Sci Eng. https://doi.org/10.1007/s13369-018-3098-1
51. Al-Betar MA, Alomari OA, Abu-Romman SM (2020) A TRIZ-inspired bat algorithm for gene selection in cancer classification. Genomics 112(1):114–126
52. Al-Betar MA, Hammouri AI, Awadallah MA, Doush IA (2020) Binary $$\beta $$ β-hill climbing optimizer with S-shape transfer function for feature selection. J Ambient Intel Hum Comput, 1–29
53. Zhang H, Sun G (2002) Feature selection using tabu search method. Pattern Recognit 35(3):701–711
54. Boughaci D, Alkhawaldeh AA (2018) Three local search-based methods for feature selection in credit scoring. Vietnam J Comput Sci 5(2):107–121
55. Marinaki M, Marinakis Y (2015) A hybridization of clonal selection algorithm with iterated local search and variable neighborhood search for the feature selection problem. Memetic Comput 7(3):181–201
56. Bermejo P, Gámez JA, Puerta JM (2011) A grasp algorithm for fast hybrid (flter-wrapper) feature subset selection in high-dimensional datasets. Pattern Recognit Lett 32(5):701–711
57. Rostami M, Berahmand K, Nasiri E, Forouzande S (2021) Review of swarm intelligence-based feature selection methods. Eng Appl Artif Intel 100:104210
58. Talbi EG (2009) Metaheuristics: from design to implementation. Wiley
59. Banka H, Dara S (2015) A hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation. Pattern Recogn Lett 52:94–100
60. Yong Z, Dun-wei G, Wan-qiu Z (2016) Feature selection of unreliable data using an improved multi-objective PSO algorithm. Neurocomputing 171:1281–1290
61. Jain I, Jain VK, Jain R (2018) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput 62:203–215
62. Yan C, Liang J, Zhao M, Zhang X, Zhang T, Li H (2019) A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. Anal Chim Acta 1080:35–42
63. Zhang T, Ding B, Zhao X, Yue Q (2018) A fast feature selection algorithm based on swarm intelligence in acoustic defect detection. IEEE Access 6:28848–28858
64. Qasim OS, Algamal ZY (2018) Feature selection using particle swarm optimization-based logistic regression model. Chemom Intell Lab Syst 182:41–46
65. Prasad Y, Biswas KK, Hanmandlu M (2018) A recursive PSO scheme for gene selection in microarray data. Appl Soft Comput 71:213–225
66. Gunasundari S, Janakiraman S, Meenambal S (2018) Multiswarm heterogeneous binary PSO using win-win approach for improved feature selection in liver and kidney disease diagnosis. Comput Med Imaging Graph 70:135–154
67. Pashaei E, Pashaei E, Aydin N (2019) Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization. Genomics 111(4):669–686
68. Xue Y, Tang T, Pang W, Liu AX (2020) Self-adaptive parameter and strategy based particle swarm optimization for large-scale feature selection problems with multiple classifiers. Appl Soft Comput 88:106031
69. Shunmugapriya P, Kanmani S (2017) A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid). Swarm Evol Comput 36:27–36
70. Zhang Y, Gong DW, Gao XZ, Tian T, Sun XY (2020) Binary differential evolution with self-learning for multi-objective feature selection. Inf Sci 507:67–85
71. Wang XH, Zhang Y, Sun XY, Wang YL, Du CH (2020) Multi-objective feature selection based on artificial bee colony: an acceleration approach with variable sample size. Appl Soft Comput 88:106041
72. Arslan S, Ozturk C (2019) Multi hive artificial bee colony programming for high dimensional symbolic regression with feature selection. Appl Soft Comput 78:515–527
73. Grover P, Chawla S (2020) Text feature space optimization using artificial bee colony. In: Das KN, Bansal JC, Deep K, Nagar AK, Pathipooranam P, Naidu RC (eds) Soft computing for problem solving. Springer, Singapore, pp 691–703
74. Hancer E, Xue B, Zhang M, Karaboga D, Akay B (2018) Pareto front feature selection based on artificial bee colony optimization. Inf Sci 422:462–479
75. Balakumar J, Mohan SV (2019) Artificial bee colony algorithm for feature selection and improved support vector machine for text classification. Inf Discov Deliv. https://doi.org/10.1108/IDD-09-2018-0045
76. Manoj RJ, Praveena MA, Vijayakumar K (2019) An ACO–ANN based feature selection algorithm for big data. Clust Comput 22(2):3953–3960
77. Janani R, Vijayarani S (2020) Automatic text classification using machine learning and optimization algorithms. Soft Comput, 1–17
78. Moradi P, Rostami M (2015) Integration of graph clustering with ant colony optimization for feature selection. Knowl-Based Syst 84:144–161
79. Dadaneh BZ, Markid HY, Zakerolhosseini A (2016) Unsupervised probabilistic feature selection using ant colony optimization. Expert Syst Appl 53:27–42
80. Liu Y, Wang Y, Ren X, Zhou H, Diao X (2019) A classification method based on feature selection for imbalanced data. IEEE Access 7:81794–81807
81. Tawhid MA, Dsouza KB (2018) Hybrid binary bat enhanced particle swarm optimization algorithm for solving feature selection problems. Appl Comput Inf. https://doi.org/10.1016/j.aci.2018.04.001
82. Liu F, Yan X, Lu Y (2020) Feature selection for image steganalysis using binary bat algorithm. IEEE Access 8:4244–4249
83. Ghanem WA, Jantan A (2019) An enhanced bat algorithm with mutation operator for numerical optimization problems. Neural Comput Appl 31(1):617–651
84. Xiang J, Han X, Duan F, Qiang Y, Xiong X, Lan Y, Chai H (2015) A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method. Appl Soft Comput 31:293–307
85. Taradeh M, Mafarja M, Heidari AA, Faris H, Aljarah I, Mirjalili S, Fujita H (2019) An evolutionary gravitational search-based feature selection. Inf Sci 497:219–239
86. Zhang Y, Gong DW, Cheng J (2015) Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM Trans Comput Biol Bioinf 14(1):64–75

87. Marie-Sainte SL, Alalyani N (2020) Firefly algorithm based feature selection for Arabic text classification. J King Saud Univ Comput Inf Sci 32(3):320–328
88. Selvakumar B, Muneeswaran K (2019) Firefly algorithm based feature selection for network intrusion detection. Comput Secur 81:148–155
89. Elyasigomari V, Lee DA, Screen HR, Shaheed MH (2017) Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. J Biomed Inform 67:11–20
90. Prabukumar M, Agilandeeswari L, Ganesan K (2019) An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. J Ambient Intell Humaniz Comput 10(1):267–293
91. Jayaraman V, Sultana HP (2019) Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. J Ambient Intel Hum Comput. https://doi.org/10.1007/s12652-019-01193-6
92. Ibrahim RA, Ewees AA, Oliva D, Abd Elaziz M, Lu S (2019) Improved salp swarm algorithm based on particle swarm optimization for feature selection. J Ambient Intell Hum Comput 10(8):3155–3169
93. Ala'M AZ, Heidari AA, Habib M, Faris H, Aljarah I, Hassonah MA (2020) Salp chain-based optimization of support vector machines and feature weighting for medical diagnostic information systems. In: Evolutionary machine learning techniques (pp. 11–34). Springer, Singapore
94. Tubishat M, Idris N, Shuib L, Abushariah MA, Mirjalili S (2020) Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection. Expert Syst Appl 145:113122
95. Hegazy AE, Makhlouf MA, El-Tawel GS (2020) Improved salp swarm algorithm for feature selection. J King Saud Univ Comput Inf Sci 32(3):335–344
96. Neggaz N, Ewees AA, Abd Elaziz M, Mafarja M (2020) Boosting salp swarm algorithm by sine cosine algorithm and disrupt operator for feature selection. Expert Syst Appl 145:113103
97. Mafarja MM, Mirjalili S (2017) Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing 260:302–312
98. Mafarja M, Mirjalili S (2018) Whale optimization approaches for wrapper feature selection. Appl Soft Comput 62:441–453
99. Nematzadeh H, Enayatifar R, Mahmud M, Akbari E (2019) Frequency based feature selection method using whale algorithm. Genomics 111(6):1946–1955
100. Emary E, Zawbaa HM, Hassanien AE (2016) Binary grey wolf optimization approaches for feature selection. Neurocomputing 172:371–381
101. Tu Q, Chen X, Liu X (2019) Multi-strategy ensemble grey wolf optimizer and its application to feature selection. Appl Soft Comput 76:16–30
102. Abdel-Basset M, El-Shahat D, El-henawy I, de Albuquerque VHC, Mirjalili S (2020) A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Syst Appl 139:112824
103. Kazemi SMR, Seied Hoseini MM, Abbasian-Naghneh S, Rahmati SHA (2014) An evolutionary-based adaptive neuro-fuzzy inference system for intelligent short-term load forecasting. Int Trans Oper Res 21(2):311–326
104. Jiang S, Chin KS, Wang L, Qu G, Tsui KL (2017) Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. Expert Syst Appl 82:216–230
105. García-Dominguez A, Galván-Tejada CE, Zanella-Calzada LA, Gamboa-Rosales H, Galván-Tejada JI, Celaya-Padilla JM, Magallanes-Quintanar R (2020) Feature selection using genetic algorithms for the generation of a recognition and classification of children activities model using environmental sound. Mobile Inf Syst. https://doi.org/10.1155/2020/8617430
106. Chiesa M, Maioli G, Colombo GI, Piacentini L (2020) GARS: genetic algorithm for the identification of a robust subset of features in high-dimensional datasets. BMC Bioinf 21(1):54
107. Storn R, Price K (1997) Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359
108. Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl-Based Syst 140:103–119
109. Hancer E (2020) A new multi-objective differential evolution approach for simultaneous clustering and feature selection. Eng Appl Artif Intel 87:103307
110. Xue Y, Jia W, Zhao X, Pang W (2018) An evolutionary computation based feature selection method for intrusion detection. Secur Commun Netw
111. Moslehi F, Haeri A (2019). An evolutionary computation-based approach for feature selection. J Ambient Intel Hum Comput, 1–13
112. Khurana A, Verma OP (2020) Novel approach with nature-inspired and ensemble techniques for optimal text classification. Multimed Tools Appl 79(33):23821–23848
113. Belazzoug M, Touahria M, Nouioua F, Brahimi M (2020) An improved sine cosine algorithm to select features for text categorization. J King Saud Univ-Comput Inf Sci 32(4):454–464
114. Saini R, Kumar P, Roy PP, Pal U (2020) Trajectory classification using feature selection by genetic algorithm. In: Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 377–388). Springer, Singapore
115. Belhadi A, Djenouri Y, Lin JCW, Cano A (2020) Trajectory outlier detection: Algorithms, taxonomies, evaluation, and open challenges. ACM Trans Manag Inf Syst (TMIS) 11(3):1–29
116. Rostami M, Berahmand K, Forouzandeh S (2021) A novel community detection based genetic algorithm for feature selection. J Big Data 8(1):1–27
117. Liu W, Xiao J, Hong M (2020) Comparison on feature selection methods for text classification. In Proceedings of the 2020 4th international conference on management engineering, software engineering and service sciences (pp. 82–86)
118. Srilakshmi V, Anuradha K, Bindu CS (2020) Optimized deep belief network and entropy-based hybrid bounding model for incremental text categorization. Int J Web Inf Syst. https://doi.org/10.1108/IJWIS-03-2020-0015
119. Thirumoorthy K, Muneeswaran K (2020) Optimal feature subset selection using hybrid binary jaya optimization algorithm for text classification. Sādhanā 45(1):1–13
120. Kyaw KS, Limsiroratana S (2020) An optimization of multi-class document classification with computational search policy. ECTI Trans Comput Inf Technol (ECTI-CIT) 14(2):149–161
121. Nourmohammadi-Khiarak J, Feizi-Derakhshi MR, Razeghi F, Mazaheri S, Zamani-Harghalani Y, Moosavi-Tayebi R (2020) New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment. Iran J Comput Sci 3(1):1–11
122. Thiyagarajan D, Shanthi N (2019) A modified multi objective heuristic for effective feature selection in text classification. Clust Comput 22(5):10625–10635
123. Lee J, Park J, Kim HC, Kim DW (2019) Competitive particle swarm optimization for multi-category text feature selection. Entropy 21(6):602

124. Khalandi S, Soleimanian Gharehchopogh F (2018) A new approach for text documents classification with invasive weed optimization and naive bayes classifier. J Adv Comput Eng Technol 4(3):167–184
125. Bai X, Gao X, Xue B (2018) Particle swarm optimization based two-stage feature selection in text mining. In: 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1–8). IEEE
126. Lu Y, Chen Y (2017) A text feature selection method based on the small world algorithm. Proc Comput Sci 107:276–284
127. Majidpour H, Soleimanian Gharehchopogh F (2018) An improved flower pollination algorithm with adaboost algorithm for feature selection in text documents classification. J Adv Comput Res 9(1):29–40
128. Allahverdipour A, Soleimanian Gharehchopogh F (2018) An improved k-nearest neighbor with crow search algorithm for feature selection in text documents classification. J Adv Comput Res 9(2):37–48
129. Maruthupandi J, Devi KV (2017) Multi-label text classification using optimised feature sets. Int J Data Mining Model Manag 9(3):237–248
130. Bidi N, Elberrichi Z (2016) Feature selection for text classification using genetic algorithms. In: 2016 8th International Conference on Modelling, Identification and Control (ICMIC) (pp. 806–810). IEEE
131. Ahmad SR, Bakar AA, Yaakub MR (2015) Metaheuristic algorithms for feature selection in sentiment analysis. In: 2015 Science and Information Conference (SAI) (pp. 222–226). IEEE
132. Too J, Abdullah AR (2021) A new and fast rival genetic algorithm for feature selection. J Supercomput 77(3):2844–2874
133. Zhong Y, Wang L, Lin M, Zhang H (2019) Discrete pigeon-inspired optimization algorithm with metropolis acceptance criterion for large-scale traveling salesman problem. Swarm Evol Comput 48:134–144
134. Duan H, Qiao P (2014) Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning. Int J Intel Comput Cybern. https://doi.org/10.1108/IJICC-02-2014-0005
135. Tian AQ, Chu SC, Pan JS, Liang Y (2020) A novel pigeon-inspired optimization based MPPT technique for PV systems. Processes 8(3):356
136. Zheng Z, Duan H, Wei C (2020) Binary pigeon-inspired optimization for quadrotor swarm formation control. In: International Conference on Swarm Intelligence (pp. 71–82). Springer, Cham
137. Pan JS, Hu P, Chu SC (2021) Binary fish migration optimization for solving unit commitment. Energy 226:120329
138. Agrawal P, Ganesh T, Mohamed AW (2021) Solving knapsack problems using a binary gaining sharing knowledge-based optimization algorithm. Compl Intel Syst pp. 1–21
139. Zhou Y, Chen X, Zhou G (2016) An improved monkey algorithm for a 0–1 knapsack problem. Appl Soft Comput 38:817–830
140. Bhattacharjee KK, Sarmah SP (2014) Shuffled frog leaping algorithm and its application to 0/1 knapsack problem. Appl Soft Comput 19:252–263
141. Xu Y, Yu H, Yan Y, Liu Y (2020) Multi-component transfer metric learning for handling unrelated source domain samples. Knowl Based Syst 203:106132
142. Wang Z, Shao YH, Wu TR (2014) Proximal parametric-margin support vector classifier and its applications. Neural Comput Appl 24(3):755–764
143. Jiao Y, Du P (2016) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant Biol 4(4):320–330
144. Ghareb AS, Bakar AA, Hamdan AR (2016) Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Syst Appl 49:31–47
145. Lee J, Yu I, Park J, Kim DW (2019) Memetic feature selection for multilabel text categorization using label frequency difference. Inf Sci 485:263–280
146. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
147. Aghdam MH, Ghasem-Aghaee N, Basiri ME (2009) Text feature selection using ant colony optimization. Expert Syst Appl 36(3):6843–6853
148. Jensen R, Shen Q (2007) Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst 15(1):73–89
149. Brownlee J (2011) Clever algorithms: nature-inspired programming recipes. Jason Brownlee
150. Brownlee J (2014) Machine learning mastery. URL: http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it
151. Rahmaniani R, Crainic TG, Gendreau M, Rei W (2017) The Benders decomposition algorithm: a literature review. Eur J Oper Res 259(3):801–817
152. Ghanem TF, Elkilani WS, Abdul-Kader HM (2015) A hybrid approach for efficient anomaly detection using metaheuristic methods. J Adv Res 6(4):609–619
153. Ghanem WA, Jantan A (2019) A new approach for intrusion detection system based on training multilayer perceptron by using enhanced Bat algorithm. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04655-2
154. Mello-Román JD, Hernández A (2020) KPLS optimization with nature-inspired metaheuristic algorithms. IEEE Access 8:157482–157492
155. Khalid R, Javaid N (2020) A survey on hyperparameters optimization algorithms of forecasting models in smart grid. Sustain Cities Soc. https://doi.org/10.1016/j.scs.2020.102275
156. Bacanin N, Bezdan T, Tuba E, Strumberger I, Tuba M (2020) Optimizing convolutional neural network hyperparameters by enhanced swarm intelligence metaheuristics. Algorithms 13(3):67
157. Drake JH, Kheiri A, Özcan E, Burke EK (2020) Recent advances in selection hyper-heuristics. Eur J Oper Res 285(2):405–428
158. Burke EK, Gendreau M, Hyde M, Kendall G, Ochoa G, Özcan E, Qu R (2013) Hyper-heuristics: a survey of the state of the art. J Oper Res Soc 64(12):1695–1724