Journal of Engineering Science and Technology
ICIST2024, Vol. 20, No. 5 (2025) 141 - 157
© School of Engineering, Taylor's University

# DISCOVERING FEATURE SELECTION IN SENTIMENT ANALYSIS FOR THE MEDICAL DOMAIN USING PARTICLE SWARM OPTIMIZATION AND MULTI-SWARM PARTICLE SWARM OPTIMIZATION ALGORITHMS

NURLAILA SYAFIRA SHAPIEI, SITI ROHAIDAH AHMAD*, NURHAFIZAH MOZIYANA MOHD YUSOP, ARNIYATI AHMAD, SITI HAJAR ZAINAL RASHID

Department of Computer Science, Faculty of Defence Science and Technology, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia
*Corresponding Author: sitirohaidah@upnm.edu.my

**Abstract**

In sentiment analysis, the high dimensionality of the features vector is a key problem that decreases the accuracy of sentiment classification in obtaining the optimum subset of features. Various techniques have been suggested for feature selection, including metaheuristic approaches, such as the multi-swarm particle swarm optimization, multi objective artificial bee colony, ant colony optimization (ACO), penguin search optimization and particle swarm optimization (PSO). These techniques have produced good results in obtaining optimal feature subsets. However, the PSO algorithm is more attractive and has received more attention from the feature selection community because of its simplicity that provides fast conversion speed and is proven to be an effective feature selection technique. Premature convergence is one of the drawbacks faced by PSO in high-dimensional feature selection. On the other hand, multi-swarm particle swarm optimization (MSPSO) is useful for selecting high-dimensional features. Nevertheless, the MSPSO limited in terms of speed convergence. This review paper presents the utilization of PSO and MSPSO as feature selection techniques in various domains, especially in sentiment analysis. This paper will also present the advantages and disadvantages of PSO as a feature selection technique. As well as the advantages of MSPSO in overcoming the weaknesses of PSO. In addition to research studies that used MSPSO as a feature selection technique these studies have also identified the potential and advantages of MSPSO.

Keywords: Feature selection, Metaheuristic algorithm, Multi-swarm particle swarm optimization, Particle swarm optimization, Sentiment analysis.

## 1. Introduction

Feature selection is a crucial and fundamental stage in sentiment analysis, particularly in datasets with a high number of dimensions. Social media text data often include noise or useless elements. Choosing the most relevant features to ensure precise outcomes is the hardest part in sentiment analysis. Excessive or unrelated characteristics in a dataset can impact the effectiveness of a classification model and lead to higher computing expenses, referred to as the curse of dimensionality [1]. Text documents that contain noise or irrelevant features use feature selection to address these issues. Therefore, enhancing machine learning algorithms is crucial as they can decrease the dimensionality of the feature space, eliminate unnecessary features, choose significant features, and enhance learning accuracy.

Sentiment Analysis is a method used to acquire, examine, and classify significant materials from social media into positive and negative categories. Textual sentiment analysis is conducted at three levels within a text: sentence, document, and feature [2]. Sentiment analysis involves five crucial steps [1]: first, text preprocessing is conducted to rectify data containing spelling errors, poor grammar, and non-existent words or concepts. Once the cleaning process is complete, the subsequent phase involves extracting pertinent features from the textual data. Feature selection is the second step in the process whereby features are evaluated and chosen to create an optimum feature subset that can accurately replicate the original feature list without altering or compromising feature quality [1]. Third, the correlation between each characteristic is determined in the ideal subset of features and the sentiment word in the textual data. The fourth step involves the identification and categorisation of emotions based on whether they exhibit positive or negative sentiments. The final stage involves the examination and assessment of the precision of feature categorisation carried out in the preceding step. The feature selection process, which is crucial for selecting pertinent and correlated features, can have significant impact on the outcomes of the conducted analysis. Hence, feature selection investigation must be conducted meticulously.

The multi-swarm particle swarm optimization (MSPSO) algorithm is a modified variant of the prevalent particle swarm optimization (PSO) algorithm comprising many sub-swarms that broadens the search space, offering more viable option in addressing challenging optimisation problems [3]. Unlike the standard PSO, exploiting a single swarm, often faces high computational complexity and suffers from an inefficient balance between exploration and exploitation, which may impact its convergence and reduce its effectiveness [4]. Consequently, MSPSO signifies to be highly effective for complex optimisation task. For more detailed explanation of the MSPSO algorithm and its implementation, the related work section provides a thorough discussion of this study's application.

Several digital libraries and databases were employed to obtain research articles for this study. This work is based on an analysis from various years of the most common articles, including citations and publications from ScienceDirect, IEEE Xplore, Springer, Wiley, Elsevier, ACM, Hindawi, MDPI and Google Scholar. The internet search for research articles have been done or figured to be an in-depth study of digital libraries using the following specific keywords, "feature selection", "metaheuristic algorithm", "particle swarm optimization", "multi-swarm particle swarm optimization", "sentiment analysis" and "medical domain".

The structure of this paper is as follows: Sections 2 offers an extensive examination and discourse on sentiment analysis, feature selection, PSO and MSPSO, with in-depth information on each topic. Section 3 explores and characterises the utilisation of sentiment analysis in the medical field, drawing on prior research. Finally, this paper is concluded with a concise overview in Sections 4 and 5.

## 2. Related Work

### 2.1. Sentiment analysis

Sentiment analysis integrates text mining, natural language processing, and artificial intelligence, which have garnered the interest of numerous academics in these domains [1, 2]. The primary difficulty in this field is enhancing the feature selection quality by identifying and eliminating irrelevant and overlapping characteristics while efficiently managing a vast feature space. Researchers are currently facing challenges in determining an appropriate method to select the optimal subset of features from the original feature space in order to decrease dimensionality and enhance the accuracy of the classification process.

### 2.2. Feature selection

A research work [5] has identified four main reasons for implementing feature selection: 1) reducing the dimensionality of feature space; 2) accelerating learning algorithms; 3) enhancing the predictive accuracy of classification algorithms; and 4) improving the interpretability of learning outcomes.

Feature selection methods are used to produce a subset of the original set of features based on certain feature selection criteria; this process will reduce calculation time and high-dimensional feature space [6-8]. In mathematical terms, feature selection is defined as follows: for classification issues, let $X^{m\times n} = \{x_{i,j}\}$ be a matrix with m features and n data samples, each of which is assigned to a certain class. The objective of feature selection is to identify the k most representative or informative characteristics (referred to as $S^{k\times n}$) from m (where $S^{k\times n} \in X^{m\times n}$ and $k << m$). These features can be considered the most discriminative for class distinction [9, 10]. A standard feature selection method consists of four fundamental components: a generation procedure to create a specific candidate subset; an evaluation function to assess the quality of the candidate subset; a stopping criterion to determine when to halt the feature selection process; and a validation procedure to verify the validity of the chosen subset [11].

Current feature selection techniques can be categorised into filter, wrapper, and embedding methods. Filter methods are used to assess the overall properties of data using variable ranking approaches to assign scores to the attributes. Features that fall below a predefined threshold can be eliminated as a consequence.

Wrapper approaches utilise a predetermined learning process to assess the effectiveness of the selected feature subsets. Embedded approaches incorporate the search for an optimal subset of features directly into the classifier generation process [9, 10]. Wrapper methods often yield superior outcomes compared to filter methods, even though they are less efficient. Wrapper approaches rely on the modelling algorithm to generate and evaluate every subset as well as on different search strategies for subset construction [12].

Chaudhuri [13] have classified search methods into three categories, exponential, sequential, and randomised selection processes. With the exponential method, the number of evaluated features will grow exponentially in relation to the size of the features. This method yields precise outcomes, but it is not feasible due to its large computing expenses. Exponential search strategies may include exhaustive search and the branch and bound method [14, 15]. Meanwhile, sequential algorithms may include linear forward selection, floating forward or backward selection, and the best first for adding or removing characteristics one after the other. Once a feature is added or removed in the chosen subset, it cannot be altered further, potentially resulting in local optima. Randomised algorithms utilise unpredictability to navigate the search space, thus preventing them from getting stuck in local optima. Randomised algorithms, such as simulated annealing, random generation, and metaheuristic algorithms, are commonly known as population-based approaches [12].

Feature selection involves solving a combinatorial optimisation issue. For a dataset with N dimensions, there are $2^N$ potential combinations of feature subsets that must be assessed [13]. Metaheuristic optimization algorithms have demonstrated their effectiveness in solving combinatorial optimisation issues [13]. Many feature selection strategies that utilise metaheuristic optimization algorithms have been proposed in previous research. The most successful and effective methods for solving various high-dimensional feature selection issues are the ant colony optimization algorithm [1], particle swarm optimization [16], genetic algorithm [17], artificial bee colony optimization [18] and firefly algorithm [19].

The PSO is more appealing to the feature selection community due to its simple algorithm, rapid processing speed, and effectiveness in feature selection processes [20]. Premature convergence is a common issue in PSO when applied to high-dimensional feature selection, while MSPSO is beneficial for selecting high-dimensional characteristics [21].

## 2.3. Feature selection in sentiment analysis in various domains

This section will briefly present feature selection methods based on sentiment analysis in various domains. Considering the availability of opinions on the web, sentiment analysis can be employed in a variety of fields [22]. Feature selection is an excellent sentiment analysis method for identifying the best subset of features while allowing researchers to choose useful and relevant characteristics that will help classifiers to produce accurate findings [23, 24]. Table 1 presents a collection of research on sentiment analysis from 2019 to 2022 in different domains, such as digital currency market (cryptocurrency), movie reviews, restaurant reviews, product reviews, and the US Airline Twitter. These projects have been conducted using various feature selection methods.

**Table 1. A summary of feature selection methods based on sentiment analysis in various domains.**

| Authors | Feature selection | Domain |
|---|---|---|
| **[25]** | Iterative semi-supervised feature selection (ISSFS) | **Digital Currency Market (cryptocurrency)** |
| **[26]** | **Binary particle swarm optimization (BPSO)** | **Movie review (imdb.com), restaurant review (yelp.com), product review** |

| | | |
|---|---|---|
| | | **(amazon.com)** |

**Table 1 (continue). A summary of feature selection methods based on sentiment analysis in various domains.**

| Authors | Feature selection | Domain |
|---|---|---|
| **[27]** | Mutual information (MI) | US Airline Twitter |
| **[28]** | Genetic algorithm (GA) Principal component analysis (PCA) Forest optimization algorithm (FOA) Particle swarm optimization (PSO) | Product review |
| **[29]** | Particle swarm optimization (PSO) | Product review (Amazon) |

## 2.4. Metaheuristic algorithm

Metaheuristic algorithms, which were discovered in 1966, are advanced optimisation methods designed to acquire optimal or near-optimal solution for complex optimisation problems [12]. Numerous metaheuristic algorithms have been developed over the years and widely used, such as genetic algorithm (GA), ant lion optimizer (ALO), particle swarm optimization (PSO), ant colony optimization (ACO), whale optimization algorithm (WOA), simulated annealing (SA) and tabu search (TS) [30, 31]. Metaheuristic methods can address different types of problems and provide effective solutions in an adequate period of time [32] and can be used for selecting a range of best features [33]. Experts have observed that metaheuristic algorithms have the potential to be applied as feature selection in sentiment analysis [30]. Numerous studies have used metaheuristic algorithms as feature selection to address issues with high dimensional datasets (HDD) [30]. A previous study has shown that metaheuristic optimisation methods can simplify optimisation, classification, and feature selection [31]. More importantly, most applications that utilise these methods have proven their effectiveness and efficiency in tackling significant and challenging problems [32]. Table 1 in Section 3.1 shows that certain metaheuristic algorithm techniques, comprising BPSO, GA, FOA and PSO, have been effectively employed for feature selection across various domains related to sentiment analysis.

### 2.4.1. Particle swarm optimization (PSO)

Kennedy developed the initial concept of particle swarm optimization (PSO) in 1995 [34, 35]. The initial idea behind the PSO was to replicate social behaviours as a model of how creatures move, for example, in a flock of birds or school of fish. In recent years, the PSO has garnered the attention of academics due to its numerous benefits and proven promise, particularly in the field of feature selection. Despite the effectiveness of the PSO, researchers are now addressing several issues associated with it, for example, controlling the associated parameters and determining the optimal setting for each iteration. Incorrect values might lead to inaccurate changes in velocity [21]. Moreover, the PSO algorithm is susceptible to becoming caught in local optima, leading to limited information exchange among particles [36]. Multiple innovative variants of the PSO have been developed with the aim of greatly enhancing the performance of optimisation processes [21]. The enhancement of the PSO involves the development of novel controlling parameter

techniques, integration of PSO with other established metaheuristic algorithms, collaboration, and implementation of multi-swarm approaches [21]. It is crucial to emphasise that there have been various PSO variations suggested in recent years.

Qiu [37] discussed the enhancements made to the traditional PSO in order to address its limitations, based on earlier research. They indicated that the traditional PSO is prone to premature convergence, resulting in sluggish convergence when dealing with high-dimensional data.

### 2.4.2. Multi-swarm particle swarm optimization (MSPSO)

The MSPSO is a better version of the standard particle PSO. The particle swarm would be divided into multiple sub-swarms, each pursuing a distinct aim to decrease information transmission among particles [21, 38]. During the early stages of the evolutionary process, the MSPSO will partition the population into several sub-swarms of equal size and tiny scale [37]. Upon conducting a thorough analysis of the application of the PSO in previous research, as mentioned in Section 3.2.1, its specific shortcomings and limitations were uncovered.

To address the limitations of PSO algorithms, researchers have developed enhance iterations such as MSPSO. For basic comparison purposes, the PSO that is typically working for a simple, robust and fast optimiser may suffer from premature convergence to local optima, whereas MSPSO one of the variants modified from the preliminary PSO algorithms, does not have that issue. The MSPSO is aimed to overcome the drawbacks of premature convergence, with just a few parameters to adjust [39, 40]. MSPSO demonstrates superior performance compared to traditional feature selection algorithms while selecting the same dimensions [41].

The study carried out by Jing [42] introduced a new MSPSO technique to tackle the problem of premature convergence. Based on evaluations completed by Qiu [37], the MSPSO exhibited higher levels of accuracy and robustness. A study conducted by Qiu [43] has demonstrated that MSPSO can effectively employ a smaller set of features while handling datasets with a large number of dimensions. Numerous research studies have primarily employed the concept of the MSPSO, thus demonstrating its effectiveness. The following section will provide a quick overview of different strategies that employ the MSPSO algorithms. Table 2 summarises 7 research studies from 2016 to 2023 on the MSPSO method for solving problems in various domains.

Qiu [37] assessed the effectiveness of a multi-swarm topology to optimise the PSO for feature selection. This method was used to divide the population into sub-swarms, followed by deploying elite strategies with local operators to improve information exchange and resource exploitation. This study pursued nine UCI datasets and two high-dimensional microarray experiments to test the performance with the k-Nearest Neighbour (KNN) algorithm. It was found that the MSPSO far exceeded six PSO-based, three wrapper, and three filter approaches in 10 of 11 datasets, with accuracy ranging from 70% to 98.53%. It performed well on high-dimensional datasets, with faster completion times.

Shen et al. [38] reviewed the effects of particle number and swarm size on the performance of the MSPSO based on the Sphere and Rastrigin functions. Performance has been assessed using mean, standard deviation, accuracy rate, and

diversity tests. The outcomes have revealed that increasing the number of particles would improve convergence precision and speed, whereas additional swarms would minimise the likelihood of local optima. However, higher particle numbers needed more computational resources, thus reducing the impact of swarm size.

Liu et al. [41] proposed a novel feature selection technique that utilised the MSPSO to ascertain the sentiment of online course reviews. The collection of reviews was evaluated using the Massive Open Online Courses (MOOC) platforms by NetEase, Inc. The MSPSO has the ability to greatly decrease the number of features used while still achieving high accuracy in recognition. The MSPSO method has outperformed other approaches in this study by achieving an overall output that exceeded 88% of Micro F-Measure.

Qiu [43] introduced a MSPSO for addressing the limitations in high-dimensional feature selection faced by the PSO. The MSPSO maintained diversity and robust global exploration by employing adaptive regrouping to promote information flow and faster convergence, while a hybrid local operator improved local exploitation. Experiments on 11 UCI datasets using KNN classification revealed that the MSPSO-A was able to outperform five PSO-based wrapper techniques by attaining perfect classification in nine datasets (73% to 98.57%) and an average accuracy of 83.35%. Thus, the MSPSO-A was very robust, with the lowest possible standard deviation. Qiu [43] intends to compare the MSPSO-A to larger datasets and observe the filter-based feature selection method in future.

Gor et al. [44] were able to optimise the MSPSO by integrating dominance rules, mesh-adaptive direct search, and regional domination techniques. By considering well-known benchmark functions, the new MSPSO has demonstrated dynamic topology support and purposeful detection strategies for preventing local optima, hence improving exploration and exploitation. This method acquired a best cost of 0.0000 and reduced the worst-case cost to 5.172 × 10^-6. Mean best costs have decreased to 13.808 × 10^-6, with elapsed times of 21.749 seconds, demonstrating significant computing efficiency.

Kumazawa et al. [45] deployed the MSPSO in an incremental model checking approach which helped improve error identification and overcome incomplete coverage problems. The authors presented novel coverage metrics, ACdiv and ACint, for evaluating search diversity and intensity. Numerical studies with the MSPSO-safe-inc showed that it would be suitable for large-scale systems and capable of finding shorter counter examples compared to baseline approaches, thus indicating a more efficient search procedure. However, the incremental method demands domain knowledge updates and fine-tuning, and its efficacy is dependent on MSPSO-safe operation with minimal computing resource consumption.

Liu et al. [46] utilised the MSPSO-based task scheduling method for load balancing that was able to reduce completion and response times in data centre task scheduling while increasing supply chain efficiency. By utilising data from the Alibaba data centre, this method was able to enhance particle swarm fitness evaluation, reduced local optima with a multi-swarm design, and elevated search efficiency using adaptive inertia weight and initialisation methods. In comparison to others, this strategy has improved the makespan and average reaction time by 39%, while maintaining the lowest load fluctuation between machines. However, its performance has shifted the virtual machine counts and the workloads needed more tuning.

**Table 2. Multi-swarm particle swarm optimization research.**

| Authors | Domain | Advantages | Limitations |
|---|---|---|---|
| **Qiu [37]** | Data mining and pattern recognition | Increased diversity. Discovered feature subsets efficiently. Able to statistically outperform other methods. Good high dimensional selection. | Not mentioned in the paper. |
| **Shen et al. [38]** | Swarm size | Outperformed the traditional PSO. Better and larger swarm sizes. Convergence improved with size. | Demanded more computational power. |
| **Liu et al. [41]** | Online course review | Minimised redundancy and preserved discriminative features. Greater recognition accuracy and robustness. | 10.83% negative feedback indicated the need for improvement. |
| **Qiu [43]** | Feature selection in Computational Intelligence | Improved classification accuracy and minimised amount features. Good exploration. Good balance in global and local searches. | Not mentioned in the paper. |
| **Gor et al. [44]** | Optimisation algorithms in Computational Intelligence | Dynamic topology and purposeful detecting strategy (PDS) support. Improved exploration and exploitation. PDS avoided local optima traps. | Lacking real-world testing. |
| **Kumazawa et al. [45]** | Model checking | Incremental coverage: Addressed state space issues. Novel metrics: Helped understand model checking. Suitable for large systems. | Complex tuning. Ran numerous times without depleting computational resources. |
| **Liu et al. [46]** | Task scheduling in sustainable supply chain data centres | Improved fitness evaluation load balancing. Adaptive design improved search efficiency and convergence speed. Avoided falling into local optima. | Complexity with elastic changes. |

According to the study by Qiu [37], the MSPSO exceeded the performance of the PSO over all 11 datasets and attained high classification accuracy in a wrapper comparison experiment.

In the summary of the literature review in Table 2, the MSPSO is successful in resolving the PSO's limitations by expanding its exploration, avoiding local optima and addressing state space problems. The dynamic topology and the PDS have also increased the search efficiency and convergence speed. The MSPSO proven to be excellent at reducing redundancy, preserving discriminate features and increasing diversity, which resulted in improved accuracy and robustness. This method stood out against the PSO in a high-dimensional selection, classification accuracy and

parameter estimation, thus making it suitable for complicated functions and substantial systems. These distinct benefits have demonstrated the superiority and competitiveness of the MSPSO in a variety of domains.

## 3.Sentiment analysis in the medical domain

The medical domain can be defined as an extensive range of health care delivery dimensions involving surgery and anaesthesiology, acute care or emergency medicine, administration, home care or self-management, primary care, intensive care, and pharmacy that vary from which study applies or what findings are being implemented [47]. This review has prioritised understanding previous sentiment analyses in medical domains to highlight its potential in order to mitigate substantial implications for patients, medical professionals, and pharmaceuticals for a variety of purposes. Sentiment analysis in the context of medical and healthcare applications is potentially a promising field for research [48].

According to Denecke and Deng [49], sentiment analysis is applicable in the domain of medicine based on several document categories which include nurse letters, radiological reports, discharge summaries, drug reviews, medblogs, and slashdot interviews. Sentiment analysis is a significant application when leveraged with other methods, as it can provide insights into patient experiences, satisfaction levels, and emotional responses by promising important information regarding patient outcomes as well as the potential for advancement in healthcare quality [50]. This review has identified previous medical studies that utilised the sentiment analysis method to depict the emotions present in health-related documents. Table 3 summarises the key findings from these studies, by highlighting the types of documents that have been analysed, the sentiment analysis techniques used, and the main outcomes.

Brief descriptions of each paper are given in the following section. Gao et al. [51] observed 28-day in-hospital mortality among sepsis patients using sentiment analysis of nursing notes, a COX model, and prognostic index (PI) from the MIMIC-III database. The results showed that higher sentiment scores were able to predict reduced mortality risks, with a PI > 0.561 signifying earlier mortality. Limitations with this method included clinician variability in MIMIC-III data, deprived generalisability, and ambiguous sentiment sources.

Guo et al. [52] have evaluated social media text data during COVID-19 by utilising logistic regression models with sparse matrices. From July 2020 to June 2021, almost 100,000 daily comments from 12 microblogs and 2,000 sample remarks were analysed, with geographic data incorporated. The machine learning logistic regression (ML-LR) model has accurately tracked shifting sentiment patterns, which suggested that public sentiments have improved during COVID-19.

Kumar et al. [53] attempted to develop a predictive model that could utilise the standard Bag of Words (BoW) method using syntactic sentiment dimensions from nursing notes obtained from the MIMIC-III v 1.4 database (2001-2012). The BoW model has accurately predicted death over varying intervals, notably within 30 days. However, longer-term predictions showed a drop in performance. Sentiment polarity scores from patient notes were an outstanding predictor of survival.

Müller and Salathé [54] advocated the development of real-time automated methods and enhanced data accuracy through ongoing crowdsourced labelling of

public social media data, mainly Twitter's API that emphasised vaccination perspectives. This method automated data collection into machine learning training by increasing public health research. The initial findings suggested the presence of generally neutral or favourable comments, with <10% of anti-vaccination.

Rao et al. [55] have built a pharmaceutical recommender system by surveying patients and then, using a selection of vectorisation methods and algorithms. Sentiment polarity was explored using the dictionary sentiment analysis, while three classifiers were tested: Gaussian Naïve Bayes, decision tree and support vector machine, with the Naïve Bayes attaining the highest accuracy.

Sasangohar et al. [56] conducted interviews with family members that were incorporating the vICU service and emphasising their experiences, issues, and recommendations for advancement. Data were gathered via phone interviews, and 86% of the participants felt satisfied, despite facing communication and technological challenges. Suggestions comprised enhancements in access, scheduling, and system capabilities. Patient communication barriers and technical issues were among the difficulties found in this study.

Bobicev and Sokolova [57] performed sentiment analysis for online medical forums, specifically in IVF discussions. These researchers utilized the multi-class and multi-label classification methods comprising Naïve Bayes (NB), support vector machine (SVM), KNN and decision trees (DT), along with information gain (IG) in feature selection. This process has accurately detected multiple states' sentiments and health concerns within a single sentence. Despite its effectiveness, this study encountered several problems due to linguistic issues and system limitations.

Waheeb et al. [58] performed sentiment analysis on medical discharge summaries using deep learning algorithms including TF-IDF, Word2Vec, GloVe, FastText, and BERT. Data from 1,237 records on obesity and 15 additional conditions were acquired and processed on the i2b2 server. These methods produced high accuracy rates, which outperformed previous advanced sentiment analysis techniques.

Weissman et al. [59] have also probed sentiment analysis methods by attending to clinical notes retrieved from MIMIC III database that composed of 793,725 ICU patient notes. To predict in-hospital mortality, the following methods were subjected to evaluation: CoreNLP, Pattern, sentimenr, Opinion, AFINN and EmoLex including logistic regression and random forest models. They found that four different methods; CoreNLP, Pattern, sentimenr and Opinion were able to determine significant associations with death risk, but lexical coverage was insufficient in the medical domain.

Table 3 lists several state document categories that have been reviewed in the present work, comprising nursing notes, clinical notes, public opinion analysis, digital health, Tele Critical Care that emphasises on interviews, medical text of discharge summaries, online medical forums, and drug reviews. This review has highlighted the numerous domains covered in these research studies, thereby presenting an extensive overview of sentiment analysis in a medical context.

A range of sentiment analysis methods, encompassing TextBlob, Valence Aware Dictionary for Sentiment Reasoner Package (VADER), dictionary sentiment polarity, manual sentiment analysis, CoreNLP, Pattern, AFINN, EmoLex and a few others have been deployed in the literature to analyse the

sentiments expressed in these documents. In this review of sentiment analysis applications in the medical domain, several studies have demonstrated the benefits and challenges associated with this approach.

**Table 3. A summary of medical studies that applied sentiment analysis methods for health-related documents.**

| Authors | Domain | Methods |
|---|---|---|
| **Gao et al. [51]** | Sentiments in nursing notes | TextBlob (sentiment polarity and subjectivity)<br>COX regression analysis |
| **Guo et al. [52]** | Public opinion analysis | Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT)<br>Machine Learning Logistic Regression (ML-LR) model with sparse matrix |
| **Kumar et al. [53]** | Health informatics (clinical notes mining) | TextBlob NLP library<br>BoW model |
| **Müller and Salathé [54]** | Public health (digital health) | Crowdbreaks :<br>Python (numpy, scipy, NLTK), Flask API, Redis, Elasticsearch, Kibana |
| **Rao et al. [55]** | Drug review | NLTK<br>Count vectorizer<br>Decision tree, Naïve Bayes, support vector machine<br>Dictionary sentiment analysis (polarity) |
| **Sasangohar et al. [56]** | Tele Critical Care | Automated analysis in Python using Valence Aware Dictionary for Sentiment Reasoner Package (Valence-based)<br>Manual sentiment analysis |
| **Bobicev and Sokolova [57]** | Online medical forums | Multi-class and multi-label classification (categorised sentiments)<br>Machine learning algorithms NB, SVM, k-NN, DT<br>Information gain |
| **Waheeb et al. [58]** | Medical text domain | TF-IDF, Word2Vec, GloVe, FastText, BERT, Clustering |
| **Weissman et al. [59]** | Clinical notes | CoreNLP, Pattern, sentimenr, Opinion, AFINN, EmoLex<br>Logistic regression<br>Random forest |

## 4. Discussion

This study sought to mainly explore numerous feature selection methods in sentiment analysis across several domains, such as bitcoin, movie reviews, product reviews, and Twitter data. BPSO, MI, PSO, and GA have generally performed well. These reviewed findings have demonstrated the robustness and easy adaptation of different feature selection methods, thus confirming their significance towards enhancing sentiment analysis results. The PSO is a prominent metaheuristic algorithm that is widely recognised for its simplicity of use and effectiveness in attempting optimisation problems. However, the PSO has crucial limitation such as preliminary convergence and difficulty in escaping local optima.

To mitigate these issues, the MSPSO was conceived by expanding on the basic features of the PSO. It incorporates different methods to improve exploration along with avoiding local optima. This review has focused on the adoption of the MSPSO across several domains to emphasise its efficacy and optimisation potential. Despite its advantages, the implementation of the MSPSO in sentiment analysis, particularly as a feature selection method, is relatively undiscovered. Only a single research study has explored the use of the MSPSO as a feature selection method in the education domain (online course reviews).

This observation indicated the demand for further research where the benefits of the MSPSO could be harnessed. This study has also emphasised the relevance of sentiment analysis in the medical domain by focusing on nine research studies that utilised numerous sentiment analysis methods on health-related documents. The results persistently showed that sentiment analysis has improved understanding and decision-making in medical applications, with the potential of shifting healthcare practices.

This sentiment analysis approach has enhanced medical treatment while also supporting public health initiatives by documenting and evaluating patient's sentiments. The list in Table 3 has also demonstrated that the potential use of metaheuristic algorithms or optimisation approaches is not extensively studied. This knowledge gap may lead to a greater opportunity for future research on discovering advanced metaheuristic algorithms aimed at enhancing sentiment analysis in the medical domain.

## 5. Conclusion

In conclusion, the present review aimed to draw attention to the significance of feature selection methods in sentiment analysis in several domains as well as its potential in medical domain. Given that both the particle swarm optimization (PSO) and the binary particle swarm optimization (BPSO) showed promising results as metaheuristic feature selection algorithms, variants of the multi-swarm particle swarm optimization (MSPSO) method should possess the ability to mitigate shortcomings of PSO. Based on the reviewed studies from Table 2, the MSPSO holds great potential for optimising and has been proven to be able to address the disadvantages posed by the PSO, thus should be applied in sentiment analysis as a feature selection method.

The sparse number of research on the deployment of the MSPSO in sentiment analysis, particularly in the medical domain calls for additional exploration. Moving ahead, our ongoing work involves a more in-depth analysis of the MSPSO's performance in real-world medical datasets. We intend to present the findings of our study in forthcoming technical paper. The application of sentiment analysis has been effective for assessing the perspectives of patients and for the improving public health. Subsequent studies should focus on developing advanced metaheuristic algorithms for feature selection with the purpose of optimising sentiment analysis within the medical domain, while paving new research pathways.

## Acknowledgement

The authors gratefully acknowledge the Fundamental Research Grant Scheme through grant no. FRGS/1/2022/ICT6/UPNM/02/1, the Ministry of Higher

Education Malaysia and Universiti Pertahanan Nasional Malaysia for supporting this research project.

## References

1. Ahmad, S.R.; Bakar, A.A.; and Yaakub, M.R. (2019). Ant colony optimization for text feature selection in sentiment analysis. *Intelligent Data Analysis*, 23, 133-158.
2. Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.
3. Lin, H.C.; Wang, P.; Lin, W.H.; and Huang, Y.H. (2021). A multiple-swarm particle swarm optimisation scheme for tracing packets back to the attack sources of botnet. *Applied Sciences*, 11(3), 1139.
4. Khan, A.; Shafi, I.; Khawaja, S.G.; de la Torre Diez, I.; Flores, M.A.L.; Galvlan, J.C., and Ashraf. I. (2023). Adaptive filtering: issues, challenges, and best-fit solutions using particle swarm optimization variants. *Sensors*, 23(18), 7710.
5. Liu, H.; and Motoda, H. (1998). *Feature extraction, construction and selection*. Springer US, Boston, MA.
6. Saraswathi, N.; Sasi Rooba, T.; and Chakaravarthi, S. (2023). Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews. *Measurement: Sensors*, 29.
7. Cai, J.; Luo, J.; Wang, S.; and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.
8. Bakshi, A.M.; and Kopparapu, S.K. (2022). Duration-normalized feature selection for Indian spoken language identification in utterance length mismatch. *Journal of Engineering Science and Technology*, 17(3), 2120-2134.
9. Tsai, C.-F.; and Sung, Y.-T. (2020). Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. *Knowledge-Based Systems*, 203, 106097.
10. Tsai, C.-F.; Chen, K.-C.; and Lin, W.-C. (2024). Feature selection and its combination with data over-sampling for multi-class imbalanced datasets. *Applied Soft Computing*, 153, 111267.
11. Dash, M.; and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1, 131-156.
12. Agrawal, P.; Abutarboush, H.F.; Ganesh, T.; and Mohamed, A.W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *IEEE Access*, 9, 26766-26791.
13. Chaudhuri, A. (2024). Search space division method for wrapper feature selection on high-dimensional data classification. *Knowledge-Based Systems* 291, 111578.
14. Sun, Z.; Bebis, G.; and Miller, R. (2004). Object detection using feature subset selection. *Pattern Recognition*, 37, 2165-2176.
15. Jain, A.K.; Duin, R.P.W.; and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37.

16. Rashno, A.; Shafipour, M.; and Fadaei, S. (2022). Particle ranking: An efficient method for multi-objective particle swarm optimization feature selection. *Knowledge-Based Systems*, 245, 108640.
17. Saibene, A.; and Gasparini, F. (2023). Genetic algorithm for feature selection of EEG heterogeneous data. *Expert System Applications*, 217, 119488.
18. Anuradha, K.; Krishna, M.V.; and Mallik, B. (2024). Bio inspired Boolean artificial bee colony based feature selection algorithm for sentiment classification. *Measurement*: *Sensors*, 32, 101034.
19. Bacanin, N.; Venkatachalam, K.; Bezdan T.; Zivkovic, M.; and Abouhawwash, M. (2023). A novel firefly algorithm approach for efficient feature selection with COVID-19 dataset. *Microprocessors and Microsystems*, 98, 104778.
20. Gad, A.G. (2022). Particle swarm optimization algorithm and its applications: A systematic review. *Archives of Computational Methods in Engineering*, 29, 2531-2561.
21. Shami, T.M.; El-Saleh, A.A.; Alswaitti, M.; Al-Tashi, Q.; Summakieh, M.A.; and Mirjalili, S. (2022). Particle swarm optimization: A comprehensive survey. *IEEE Access*, 10, 10031-10061.
22. Alassaf, M.; and Qamar, A.M. (2020). Aspect-based sentiment analysis of Arabic tweets in the education sector using a hybrid feature selection method. *Proceeding of the* 14th *International Conference on Innovations in Information Technology* (*IIT*), Al Ain, United Arab Emirates.
23. Hung, L.P.; Alfred, R.; and Hijazi, M.H.A. (2015). A review on feature selection methods for sentiment analysis. *Advanced Science Letters*, 21, 2952-2956.
24. Ighazran, H.; Alaoui, L.; and Boujiha, T. (2018). Metaheuristic and evolutionary methods for feature selection in sentiment analysis (a comparative study). *Proceedings of the* 2018 *International Symposium on Advanced Electrical and Communication Technologies* (*ISAECT*), Rabat, Morocco, 1-6.
25. Akba, F.; Medeni, I.T.; Guzel, M.S.; and Askerzade, I. (2020). Assessment of iterative semi-supervised feature selection learning for sentiment analyses: Digital currency markets. *Proceedings of the* 2020 *IEEE* 14th *IEEE International Conference on Semantic Computing* (*ICSC*), San Diego, CA, USA, 459-463.
26. Botchway, R.K.; Yadav, V.; Kominkova, Z.O.; and Senkerik, R. (2022). Text-based feature selection using binary particle swarm optimization for sentiment analysis. *Proceedings of the International Conference on Electrical, Computer, and Energy Technologies* (*ICECET*), Prague, Czech Republic.
27. Utama, H. (2019). Sentiment analysis in airline tweets using mutual information for feature selection. *Proceedings of the* 2019 4th *International Conference on Information Technology*, *Information Systems and Electrical Engineering* (*ICITISEE*), Yogyakarta, Indonesia, 295-300.
28. Senthilkumar, V.; and Kumar, B.V. (2021). A survey on feature selection method for product review. *Proceedings International Conference on Advancements in Electrical*, *Electronics*, *Communication*, *Computing and Automation*, (*ICAECA*), Coimbatore, India.

29. Vasudevan, P.; and Kaliyamurthie, K. (2021). Product sentiment analysis using particle swarm optimization based feature selection in a large-scale cloud. *Proceedings of the First International Conference on Computing, Communication and Control System*, *I3CAC* 2021, 7-8 *June* 2021, *Bharath University*, *Chennai,* India.
30. Almufti, S.M.; Shaban A.A.; Ali, Z.A.; Ali, R.I.; and Fuente , J.D. (2023). Overview of metaheuristic algorithms. *Polaris Global Journal of Scholarly Research and Trends*, 2(2),10-32.
31. Yab, L.Y.; Wahid, N.; and Hamid, R.A. (2022). A meta-analysis survey on the usage of meta-heuristic algorithms for feature selection on high-dimensional datasets. *IEEE Access*, 10, 122832-122856.
32. Ahmad, S.R.; Yusop, N.M.M.; Asri, A.M.; and Amran, M.F.M. (2021). A review of feature selection algorithms in sentiment analysis for drug reviews. *International Journal of Advanced Computer Science and Applications*, 12(12), 126-132.
33. Ahmad, S.R.; Bakar, A.A.; and Yaakub, M.R. (2015). Metaheuristic algorithms for feature selection in sentiment analysis. *Proceedings of the* 2015 *Science of Information Conference* (*SAI*), London, United Kingdom, 222-226.
34. Kennedy, J.; and Eberhart, R. (1995). Particle swarm optimization. Proceedings *of the ICNN'*95*-International Conference on Neural Networks*, Perth, WA, Australia, 1942-1948.
35. Gali, V.; Gupta, N.; and Gupta, R.A. (2018). Enhanced particle swarm optimization based dc-link voltage control algorithm for interleaved SAPF. *Journal of Engineering and Technology*, 13(10), 3393-3418.
36. Du, J.; Zhang, A.; Guo, Z.; Yang, M.; Li, M.; and Xiong, S. (2018). Atomic cluster structures, phase stability and physicochemical properties of binary Mg-X (X= Ag, Al, Ba, Ca, Gd, Sn, Y and Zn) alloys from ab-initio calculations. *Intermetallics*, 95, 119-129.
37. Qiu, C. (2019). A novel multi-swarm particle swarm optimization for feature selection. *Genet Program Evolvable Machines*, 20, 503-529.
38. Shen, Y.; Li, Y.; Kang, H.; Zhang, Y.; Sun, X.; Chen, Q.; Peng, J.; and Wang, H. (2018). Research on swarm size of multi-swarm particle swarm optimization algorithm. *Proceedings of the* 2018 *IEEE* 4th *International Conference on Computer and Communications* (*ICCC*), Chengdu, China, 2243-2247.
39. Freitas, D.; Lopes, L.G.; and Morgado-Dias, F. (2020). Particle swarm optimisation: A historical review up to the current developments. *Entropy*, 22(3), 362.
40. Chrouta, J.; Farhani, F.; and Zaafouri, A. (2021). A modified multi swarm particle swarm optimization algorithm using an adaptive factor selection strategy. *Transactions of the Institute of Measurement and Control*, 01423312211029509.
41. Liu, Z.; Liu, S.; Liu, L.; Sun, J.; Peng, X.; and Wang, T. (2016). Sentiment recognition of online course reviews using multi-swarm optimization-based selected features. *Neurocomputing*, 185, 11-20.
42. Jie, J.; Wang, W.; Liu, C.; and Hou, B. (2010). Multi-swarm particle swarm optimization based on mixed search behavior. *Proceedings of the 2010 5th*

*IEEE Conference on* Industrial *Electronics and Applications*, Taichung, Taiwan, 605-610.

43. Qiu, C. (2020). A multi-swarm particle swarm optimization with an adaptive regrouping strategy for feature selection. *Proceedings of the* 2020 4th *International Conference on Robotics and Automation Sciences* (*ICRAS*), Wuhan, China, 130-136.
44. Gor, I.; Gunel, K.; and Isman, G. (2023). Incorporating the regional dominance strategy into multi-swarm PSO. *Proceedings of the* 2023 5th *International Conference on Problems of Cybernetics and Informatics* (*PCI*), Baku, Azerbaijan, 1-3.
45. Kumazawa, T.; Takimoto, M.; Kodama, Y.; and Kambayashi, Y. (2023). *Enhancing safety checking coverage with multi-swarm particle swarm optimization*. In Mathieu, P.; Dignum, F.; Novais, P.; and De La Prieta, F. (Eds.), *Advances in practical applications of agents, multi-agent systems, and cognitive mimetics. The PAAMS collection*. Springer Nature Switzerland, 137-148.
46. Liu, Q.; Zeng, L.; Bilal, M.; Song, H.; Liu, X.; Zhang, Y.; and Cao, X. (2023). A multi-swarm PSO approach to large-scale task scheduling in a sustainable supply chain datacenter. *IEEE Transactions on Green Communications and Networking*, 7(4), 1667-1677.
47. Valdez, R.S.; McGuire, K.M.; and Rivera, A.J. (2017). Qualitative ergonomics/human factors research in health care: Current state and future directions. *Applied Ergonomics*, 62, 43-71.
48. Ramírez-Tinoco, F.J.; Alor-Hernández, G.; Sánchez-Cervantes, J.L.; Sala, Zarate, M.D.P.; and Valencia-Garcia, R. (2019). *Use of sentiment analysis techniques in healthcare domain*. In Alor-Hernández, G.; Sánchez-Cervantes, J.L.; Rodríguez-González, A.; and Valencia-García, R. (Eds.), *Current trends in semantic web technologies: theory and practice*. Springer International Publishing, 189-212.
49. Denecke, K.; and Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64, 17-27.
50. Usman, M.; Mujahid, M.; Rustam, F.; Flores, E.; Mazon, J.L.V.; de la Torres Diez, I.; and Ashraf, I. (2024). Analyzing patients satisfaction level for medical services using twitter data. *PeerJ Computer Science*, 10, e1697.
51. Gao, Q.; Wang, D.; Sun, P.; Luan, X.; and Wang, W. (2021). Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the MIMIC-III database. *Computational and Mathematical Methods in Medicine*, 2021, 1-9.
52. Guo, F.; Liu, Z.; Lu, Q.; Ji, S.; and Zhang, C. (2024). Public opinion about Covid-19 on a microblog platform in China: topic modeling and multidimensional sentiment analysis of social media. *Journal of Medical Internet Research*, 26, e47508.
53. Kumar, V.; Bajpai, R.; and Roy, R.B. (2022). Clinical notes mining for post discharge mortality prediction. *IETE Technical Review*, 39, 953-959.
54. Müller, M.M.; and Salathé, M.; (2019). Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in Public Health*, 7.
55. Rao, K.K.; Kona Sravya, K.; Sai, K.J.P.; Giri, G.; Saib, R.; and Ganesanc, G. (2022). Machine learning based drug recommendation from sentiment analysis

of drug rating and reviews. *Proceedings of the workshop on artificial intelligence* (*WAI* 2022) *co-located with computing congress* (*CC* 2022). Chennai, India, 3146.

56. Sasangohar, F.; Dhala, A.; Zheng, F.; Ahmadi, N.; Kash, B.; and Masud, F. (2021). Use of telecritical care for family visitation to ICU during the COVID-19 pandemic: An interview study and sentiment analysis. *BMJ Quality and Safety*, 30(9), 715-721.
57. Bobicev, V.; and Sokolova, M. (2018). Thumbs up and down: Sentiment analysis of medical online forums. *Proceedings of the* 2018 *EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, Brussels, Belgium, 22-26.
58. Waheeb, S.A.; Khan, N.A.; and Shang, X. (2022). An efficient sentiment analysis based deep learning classification model to evaluate treatment quality. *Malaysian Journal of Computer Science*, 35, 1-20.
59. Weissman, G.E.; Ungar, L.H.; Harhay, M.O.; Ahmadi, N.; Kash, B.; and Masud, F. (2019). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics*, 89, 114-121.