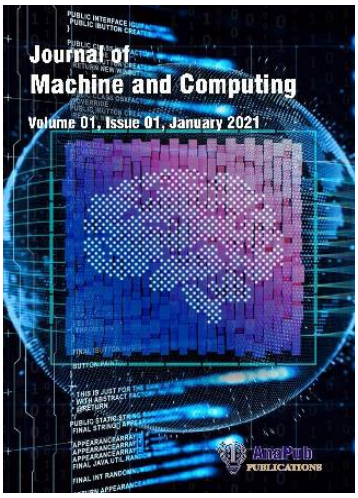
Journal Pre-proof

Enhanced Opinion Mining from Medical Tweets using an Optimized Penguin Search-Based Feature Selection Algorithm



Anuprathibha T, Pravin Kumar M, Sakthi G and Rajkumar KK

DOI: 10.53759/7669/jmc202505093
Reference: JMC202505093
Journal: Journal of Machine and Computing.

Received 01 September 2024
Revised form 18 December 2024
Accepted 16 March 2025

Please cite this article as: Anuprathibha T, Pravin Kumar M, Sakthi G and Rajkumar KK, "Enhanced Opinion Mining from Medical Tweets using an Optimized Penguin Search-Based Feature Selection Algorithm", Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505093.

This PDF file contains an article that has undergone certain improvements after acceptance. These enhancements include the addition of a cover page, metadata, and formatting changes aimed at enhancing readability. However, it is important to note that this version is not considered the final authoritative version of the article.

Prior to its official publication, this version will undergo further stages of refinement, such as copyediting, typesetting, and comprehensive review. These processes are implemented to ensure the article's final form is of the highest quality. The purpose of sharing this version is to offer early visibility of the article's content to readers.

Please be aware that throughout the production process, it is possible that errors or discrepancies may be identified, which could impact the content. Additionally, all legal disclaimers applicable to the journal remain in effect.

© 2025 Published by AnaPub Publications.

AnaPub
PUBLICATIONS

Enhanced Opinion Mining from Medical Tweets Using an Optimized Penguin Search-Based Feature Selection Algorithm

[1]T.Anuprathibha, [2]M.Pravin Kumar, [3]G. Sakthi, [4]Rajkumar KK

Associate Professor, Department of Information Technology, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Professor, Department of Medical Electronics, Velalar College of Engineering and Technology, Erode, India.
Professor, School of Computer Science & Engineering, Galgotias University, Uttar Pradesh, India
Assistant Professor, Department of Electronics and communication engineering, SNS college of Engineering, Coimbatore, Tamilnadu, India

anuprathiba16@gmail.com, pravinkumarm@velalarengg.ac.in, sakthihit@gmail.com, kkrajkumarece@gmail.com

Corresponding Author: T.Anuprathibha

ABSTRACT- Opinion mining is the approach of utilizing Natural Language Processing (NLP) concepts to extract the public opinions on specific topics and has gained increasing significance in major text mining applications. Many opinion mining methods have been developed that builds a model to collect and analyse the opinions on topics from the blogs, reviews, comments or tweets. Recently, the application of opinion mining on medical tweets has gained immense research interest due to the challenge of processing unique medical terms in tweets. In this paper, an opinion mining framework has been developed to provide automatic extraction of opinions from medical tweets using improved optimization algorithms. The input tweets undergo pre-processing, and features are extracted by POS tagging and n-grams. Then the feature subset candidates are selected using Penguin Search Optimization algorithm (PeSOA) and Improved PeSOA. In PeSOA, the solution search operation is random and does not utilize exploration concept effectively in order to maintain simplicity. The Improved PeSOA exploits this limitation and introduces a new solution search equation to compliment the traditional search process and an effective feature subset ranking concept. These concepts of Improved PeSOA increase the efficiency of selecting optimal feature subsets. Once the features are selected, the final classification is performed using k-Nearest Neighbor (k-NN), Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers to obtain the opinions. Experiments are conducted on medical datasets containing Cancer and drug tweets. The results prove that the classification accuracy for opinion mining has been increased significantly by the use of Improved PeSOA than the traditional PeSOA.

Keywords: Twitter, Opinion mining, Natural Language Processing, Naïve Bayes (NB), Penguin Search Optimization algorithm, Improved PeSOA, k-Nearest Neighbor (KNN), Support Vector Machine (SVM).

1. INTRODUCTION

Sentiment analysis and opinion mining is the field of study that examines the peoples' opinions and views towards different topics on products, services, organizations, individuals, issues and events [1]. Both sentiment analysis and opinion mining is the same field of study but some academic researchers provide distinct meanings to these terms using the linguistics. They define opinion mining as extraction of opinions of users and the sentiment analysis as extraction of emotion of users. However, these two terms are often considered as single process. Similarly, there are many names representing the opinion mining with a slightly different task. These tasks include opinion extraction, opinion mining, sentiment extraction, sentiment mining, subjectivity analysis, emotion analysis, etc. these tasks are grouped together as sentiment analysis or opinion mining [2]. Both the sentiment analysis and opinion mining terms are flexibly used in academic research works [3]. This work uses the term opinion mining as the primary term for representation of the research work.

Opinion mining combines the natural language processing and text mining applications and employs techniques like machine learning for analysing and classifying the text as positive or negative. First the opinion mining tool or application collects the text about the specified topic from various sources or particular source specified by the developers. The sources include blogs, tweets, posts, comments, reviews and messages from various interaction sites or social media sites [4], [5]. Then the text data are processed and analysed for detecting the opinion words or sentiment features. Based on these words or features, the tweet data are classified into categories of positive, negative or neutral. Opinion mining helps the people in understanding the opinion of certain individuals or group of people on an individual, product, service, event, issue or topics [6]. The opinion mining techniques are also commonly used by many organizations and service providers to find the users' exact state of mind regarding their products and services and to use them to improve their yield quality to enhance customers' satisfaction. Many organizations apply automated opinion mining to evaluate customers' sentiments and improve decision making process [7].

Authors Pre-Proof

For automated opinion mining, various approaches have been employed namely NLP, text mining, machine learning techniques like maximum entropy, NB, k-NN, SVM, neural networks (NN), decision tree algorithms, etc. [8], [9]. These algorithms were utilized in combination with feature selection methodologies to determine the sentiment polarity of the reviews and opinions. However there are various challenges in automated opinion mining. The word meaning challenge is the most common challenge in automated opinion classification as some words have different meanings based on their position on a sentence. For example, the word "small" can be used as positive term when describing the size of components as well as negative when used to describe the height of an individual. Likewise the problem of categorization of terms based on class labels is also a challenging task due to the utilization of different sets of features [10]. Many research works have been trying to overcome these challenges more effectively using novel feature selection and classification strategies [11]. However, there is another challenge that the strategies developed for particular domains are less effective in other domains. The medical domain is one such domain which requires specialized approaches to improve opinion mining as these results will be employed in various real-world applications of sensitive medical field [12].

Extracting opinions from medical tweets is considered as a difficult process as the uncommon medical terms pose greater challenge [13]. Additionally, the positive/negative sentiments of drug terms are dual-edged and hence more careful approaches are required to achieve highly accurate sentiment classification. These accurate results helps in applications like patient surveillance, tracking the patient activities on social media and analysing the psychological effects on patients regarding the illness and corresponding treatments. Tweets are most common source for medical opinion mining due to their small message length and easy access for progressive researches. Even opinion mining from tweet data also possesses many challenges. The handling of informal texts, meaningless expressions, similes and duplicate tweets are the forefront issues [14]. In this paper, an effective opinion mining framework is proposed for automated opinion extraction from medical tweets by considering all the common challenges.

The proposed approach utilizes three machine leaning algorithms in SVM, NB and k-NN algorithms for the sentiment classification process and an improved optimization algorithm for feature selection. The major contribution is the development of an improved PeSOA algorithm for the feature selection process. The traditional PeSOA algorithm is based on the food search process of penguin gang. The optimal penguin group (feature subset) with the most abundant food source is identified as the superior option. This system relies solely on search operations, minimizing the exploitation notion for simplicity, and the ranking of feature subsets is ineffective. This paper presents an enhanced PeSOA designed to address these constraints through an efficient solution search procedure and the ranking of feature subsets based on the information gain parameter. The features chosen by the enhanced PeSOA are employed by the classifiers to categorize the sentiments of the tweets. The experiments are performed to assess the efficacy of the proposed method for opinion mining. The subsequent sections of the paper are structured as follows: Section 2 examines the cutting-edge methodologies pertinent to this investigation. Section 3 delineates the proposed opinion mining methodology. Section 4 delineates the experimental findings and analyses pertaining to the suggested methodology. The paper's conclusion is presented in section 5.

2. RELATED WORKS

In recent years, many techniques have been developed for the automated opinion mining and corresponding applications. Many researches focused on developing sentiment analysis approaches using metaheuristic optimization algorithms for feature selection and machine learning algorithms for sentiment classification especially for medical related tweets. Rathan et al, [15] presented an attribute based SVM model for Twitter opinion mining with an accuracy of 86%. However, the manual creation of ontology has increased the time consumption. Ghiassi et al, [16] presented a domain transferable lexicon set and supervised machine learning approach of dynamic NN and SVM. This approach reduces the overall feature subsets and increases the sentiment classification accuracy. However this approach is not comprehensive in spam tweet removal that reduces the performance significance. Saleena, [17] introduced an ensemble classification system for twitter sentiment analysis in which the NB, RF, SVM, and LR classifiers are combined to improve the sentiment classification performance. The major limitation of this ensemble classifier is that it fails to effectively classify the neutral tweets.

Na et al, [18] proposed a rule-based linguistic approach for sentiment classification of drug reviews. This approach provided greater advantage for the drug review data handling and increased the sentiment classification accuracy to 78%. Sharif et al, [19] proposed a sentiment classification framework for detecting adverse drug reactions (ADR) with n-grams feature extraction and selection process. This approach provides an accuracy of 78.2% due to the effective feature subset representation with high discriminatory potential.

Authors Pre-Proof

Korkontzelos et al, [20] analysed the effect of sentiment analysis on ADR from tweets and forum posts using a specialized classification approach. The sentiment bearing features of ADR has increased the sentiment analysis but the non-selection of informative features results in lower accuracy.

Luna-Aveiga et al, [21] presented sentiment polarity detection approach for asthma disease management from tweet messages. This approach uses Senti-WordNet and n-grams method to identify the sentiment polarities with precision of 82.95%. However, the detection of sarcasm and irony tweets is only less efficient in these two approaches. Rodrigues et al, [22] presented a SentiHealth-Cancer tool for detecting mood of cancer patients in Twitter. This tool identified the cancer patient emotions in Portuguese tweets using n-grams and achieved an accuracy of 71.25%. Crannell et al, [23] developed a regular expression software pattern matching to filter the tweets and categorize them into appropriate sentiment labels for identifying the sentiments of US cancer-patient tweets. However this approach employs only the expression based matching while the cancer related features are not considered for classification. Salas-Zárate et al, [24] proposed a feature based sentiment analysis approach on tweets about diabetes. The approach utilized n-grams method to achieve 81.93% precision of sentiment classification but this approach is less effective in handling other health tweets.

Optimization algorithms have a significantly larger role in sentiment analysis. GA and PSO are the most common optimization algorithms employed for various applications. Kalaivani et al, [25] proposed a feature reduction technique based on information gain and GA for enhanced opinion mining. Keshavarz et al, [26] presented an adaptive lexicon learning approach using GA for solving the non-convex optimization problem of sentiment analysis in microblogs data. Onan et al, [27] proposed a feature selection model based on genetic rank aggregation for improving the sentiment classification accuracy to 94.71%. This ensemble model utilizes the feature lists obtained from many feature selection methods and employs GA to aggregate 60% of most informative features from these lists to increase the classification accuracy. Iqbal et al, [28] also presented a sentiment analysis framework using GA based feature reduction in which the GA has increased the accuracy of machine learning classifiers by 4%. However, the convergence speed of GA is much slower than other advanced optimization algorithms and also the computation and time complexity is high for these GA based feature reduction/selection approaches.

Likewise, Basari et al, [29] applied PSO algorithm for sentiment feature selection and SVM for classification. Similarly, Gupta et al, [30] also employed PSO for feature selection and conditional random fields (CRF) for classification. Akhtar et al, [31] presented a two-step sentiment analysis method using PSO feature selection and ensemble classification. This ensemble classifier combines maximum entropy, SVM and CRF to provide sentiment classification with high accuracy of 80%. However these feature selection techniques using PSO is only single objective and hence does not support multi-objective problems. Akhtar et al, [32] has also presented another feature selection approach using multi-objective optimization to overcome this limitation. Nagarajan et al, [33] proposed a hybrid sentiment analysis approach to classify the streaming Twitter data. This hybrid approach consists of GA, PSO and decision tree classifier to obtain 90% accuracy of sentiment classification. However, the computation time is high for these multi-objective PSO, optimized CNN, and hybrid approach.

Many other recent and advanced optimization algorithms have also been applied for the sentiment analysis problem. Kumar et al, [34] proposed the use of firefly algorithm for feature selection in sentiment analysis approach and increased the classification accuracy of SVM by 5.64% than other models while also supporting multiple languages. Pandey et al, [35] developed a sentiment analysis approach using hybrid cuckoo search method that combines the k-means algorithm with cuckoo search algorithm for clustering the sentiment contents with high accuracy. But this approach is not efficient in handling sarcasm and irony tweets. Alarifi et al, [36] introduced a big data sentiment analysis approach for low error rate classification which utilizes greedy algorithm for feature selection and cat swarm optimization-based long short-term memory neural networks for classification. Though efficient with high accuracy and less errors, this approach has higher text noise that degrades the overall performance. Tubishat et al. [37] improved Arabic tweet sentiment analysis using whale optimization algorithm-based feature selection. This method reduced features using information gain and classified with SVM with high accuracy. Kumar et al, [38] demonstrated the use of two swarm intelligence algorithms namely binary grey wolf and binary moth flame based optimal feature selection approaches in sentiment analysis. These approaches reduced the features by 30% while increased the sentiment classification accuracy by 10%. Du et al, [39] proposed the optimization based machine learning based approach for sentiment analysis on HPV vaccines related tweets. This approach utilized POS tags and classified using SVM and hierarchical classification with a parameter based optimization of SVM. However, this approach has low performance due to inefficient handling of unbalanced tweet data. The limitations of the state-of-the-art methods discussed in this section are considered while developing the proposed opinion mining framework in order to avert or minimize these known disadvantages.

Authors Pre-Proof

3. METHODS

The proposed opinion mining approach attempts to improve sentiment analysis of medical tweets. Pre-processing, feature extraction, selection, and classification determine tweet sentiment in the proposed method. The framework's detailed architectural diagram is shown in Figure 1. The proposed methodology employs the Twitter API to gather data on certain subjects for input purposes. The data receives pre-processing, followed by the extraction of features using feature descriptors. The properties are subsequently picked employing the PeSOA and Improved PeSOA techniques. Opinion mining uses three classifiers to evaluate classification accuracy.

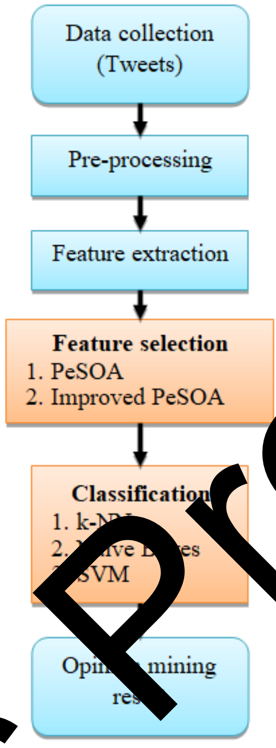


Fig.1. Architecture of proposed opinion mining framework

3.1. Data collection

Twitter API keywords related to cancer and pharmaceuticals provide the input data. Medications were mentioned in 500 of 6,000 cancer tweets. Test tweets are used after 2,500 tweets for training. We can add tweets for testing without restriction with the proposed method.

3.2. Pre-processing

Pre-processing is performed to remove the unnecessary words and irrelevant tweets in the collected datasets [40]. The pre-processing in this work consists of the following steps: data cleaning, spell check, punctuation check, URLs check, case normalization, stemming and stop word removal. Fig.2 shows the processes involved in pre-processing stage.

Authors Pre-Proof

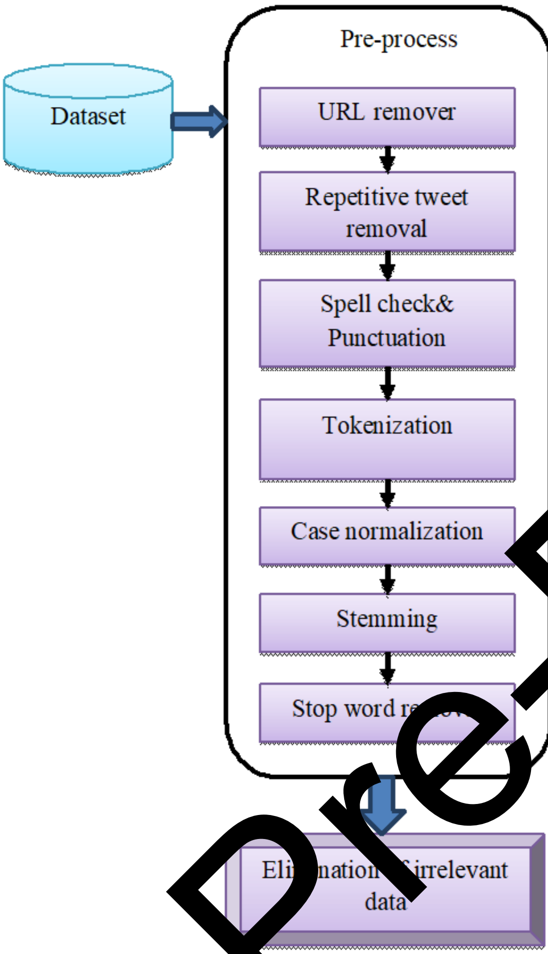


Fig.2. Pre-processing steps

The data cleaning and filtering process is the main task in pre-processing that aims to minimize the errors in data and also to reduce the noise levels. First the URLs in the tweet messages are analysed as the character limit in Twitter has provided the users to utilize the URL shortening services to minimize the content. While the shortened URLs redirect to the original end URL, the original URLs has to be checked to verify the data. This process is performed at the API level which removes the comments, links, advertisements and other irrelevant parts in a tweet. Also the repetitive tweets are also eliminated.

The data cleaning process further includes the process of spell checking using WordNet like dictionaries and punctuation checking to minimize the errors in opinion extraction. The message length detection is performed to check whether the tweet message is a single part message or multi-part message. In multi-part messages, the opinions in some parts may differ due to the use of different sentiment words in describing the same incident or topic. So the length of the messages is detected and the continuation messages are often avoided. The tokenization of the tweets is performed to replace the sensitive tweets with unique identification symbols to utilize all the information without violating security. Though the tweets are case sensitive, the detection of opinions may find difficult to handle case variations; so the cases are normalized. Finally the stemming and stop words are removed. Stemming is the process of removing '-ing' and similar prefix-suffixes that does not provide any meaning. Similarly, the stop words are the words in messages that have no individual meaning and do not impact the opinions of the messages.

3.3. Feature extraction

Feature extraction is the technique to minimize the number of aspects required to describe a dataset. If the system processes a complete dataset without aspects or features, either the system fails to process or takes long duration to complete the processing. Both these outcomes are degradable to the system efficiency; the feature extraction concept has been introduced. With feature extraction, even the complex datasets can be described by a few aspects or properties and the classification system detects and follows such aspects to

Authors Pre-Proof

categorize them. Different datasets utilize different features for increasing their classification accuracy and minimizing the processing time. In this work, the content words, function words, POS tags and POS n-grams features are extracted to improve the classifier performance [41].

Content Words – Content words are defined as words that provide independent meaning when utilized in a phrase. The majority of nouns and their defining terms has independent meanings in general.

Function Words – Function words are terms that possess minimal or uncertain meaning. These terms solely denote grammatical links among words, lacking independent meaning when viewed in isolation.

Part of speech tags – POS tags is a method of annotating a word in a tweet with reference to a corpus, identifying it as relating to a specific part of speech, based on its definition and context. This work employs parts of speech tags such as nouns, verbs, pronouns, adverbs, adjectives, and articles.

Part of speech n-grams – An n-gram model is characterized as a probabilistic language model utilized for forecasting the subsequent item in a sequence, structured as a (n - 1) order Markov model. The selected n-grams may consist of unigrams, bigrams, trigrams, or higher-order combinations, but must offer contextual relevance. This study employs trigrams, as four-grams and higher n-grams have not shown enhanced categorization in past research.

In this work, these features are utilized individually as well as in combined states. The combinations tried in this work are content words + function words, function words + POS n-grams, and content words + function words + POS n-grams. The combination features are utilized as single features in order to capture both the style and topic based aspects of the tweets. Fig.3 shows the types of features extracted in this proposed approach.

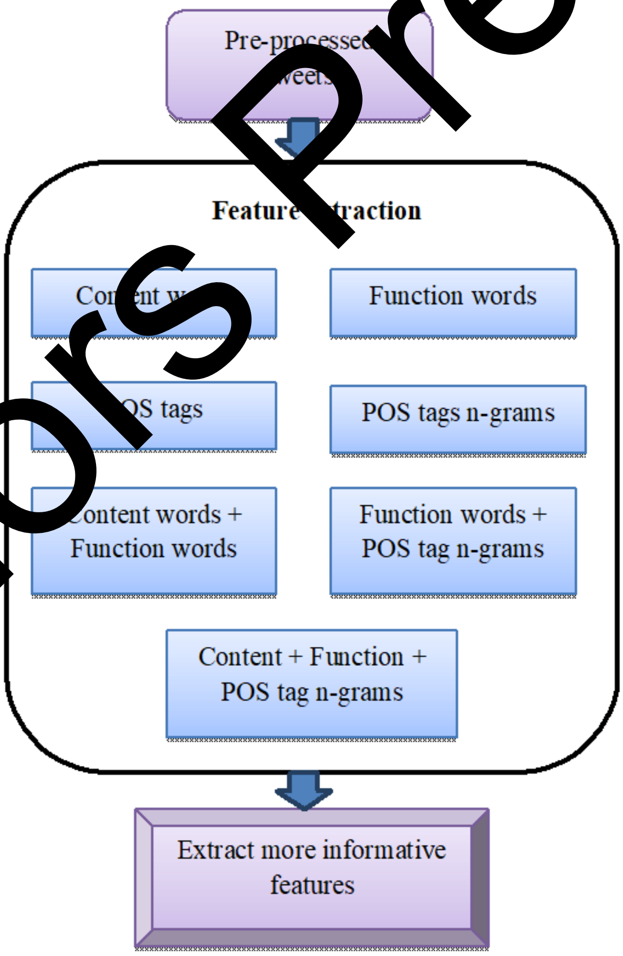


Fig.3. Feature extraction process

Authors Pre-Proof

3.4. Feature selection

Feature selection is the process of identifying one or more features that yield optimal results. In any classification application, the primary stage is to pre-assess the optimal and ideal attributes. Nonetheless, the optimal features can be discerned solely after their implementation in the classifier, a process that requires an extended duration. Thus, the feature selection issue is conceptualized as a standard problem and addressed using several methodologies to identify the optimal characteristics. A multitude of research studies have utilized ranking models for this objective. The current concept is to formulate the feature selection issue as an optimization problem and address it with sophisticated optimization methods. This study employs PeSOA and Improved PeSOA for feature selection. The PeSOA employs a conventional penguin food search approach for selection. Due to the inadequate execution of the exploration property, an enhanced PeSOA is presented in this article. The enhanced PeSOA first employs a novel solution search equation to augment the exploration notion. The features are subsequently sorted utilizing the information gain measure to facilitate reduction, followed by the selection of the optimal feature subset.

3.4.1. PeSOA feature selection:

PeSOA has been inspired by the hunting behaviour of penguins for searching the fish in ice holes [42]. The penguins have to swim deeper to harvest the fishes and hence the oxygen levels are also necessarily monitored. In this hunting process, each penguin has to search food and share their locations with the whole group. Then all the locations are analysed and the location with high amount of food is chosen by the whole group to make a move to that location for hunting.

Initially, the entire penguin society is segmented into many sorts of groupings, each of which navigates towards the fish location randomly. If the food supply is inadequate, the group relocates to new areas. The initial movement relies on random solutions, allowing the penguin groups to select their own hunting locations. In this study, penguins are selected as the characteristics, and the groups are regarded as subsets of features. Therefore, the optimal feature subsets are those penguins with the most advantageous food locations. A random population of P solutions (features) is generated. This movement is expressed as

$$X_{new} = X_{old} + rand * (X_{l\,best} - X_{l\,old}) \quad (1)$$

Where X_{new} is the new solution, X_{old} is the old solution. The overall processes in PeSOA for feature selection are provided in the following pseudo code.

Figure 4 shows the flowchart which represents an optimization algorithm inspired by penguins. It begins by initializing M penguins and their positions. The positions are updated iteratively using an equation until a termination condition is met. If $R/2>0$, the global best (gbest) and best individual positions (xbest) are updated. Finally, the algorithm outputs the best global solution before stopping.

Authors Pre-Proof

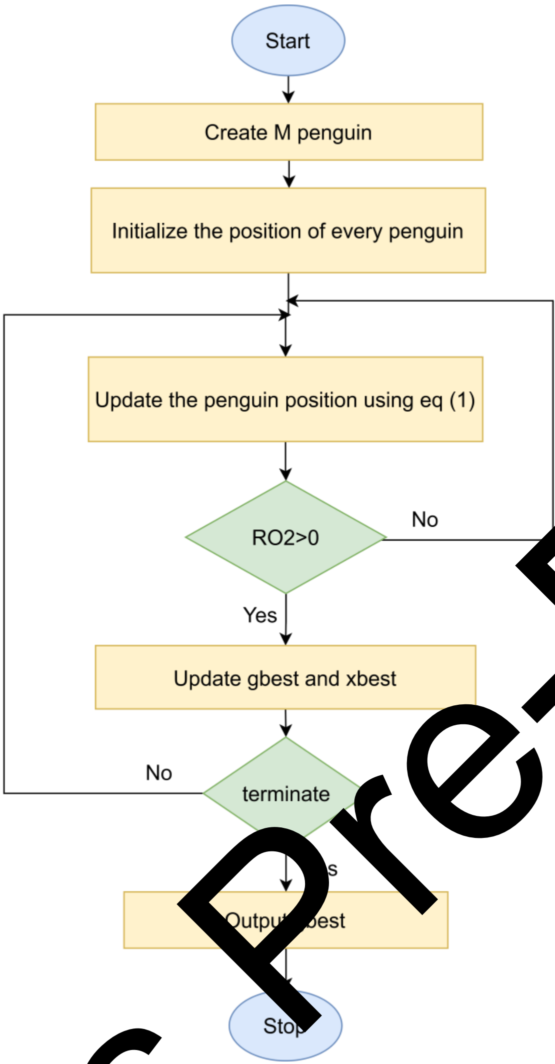


Fig 4. Flowchart of PeSOA

3.4.2. Improved PeSOA feature selection:

In the PeSOA, the random step search of the penguins is not effective for capable exploration. Hence a new solution search process is initiated. First the population is randomly generated and the initial solution n_i can be formulated using

$$n_i = n_{min} + rand(0,1) * (n_{max} - n_{min}) \quad (2)$$

Where $i \in (1,2, \dots N)$, n_{max} and n_{min} are the lower and upper bounds of n_i. This initial solution is based on the minimum and maximum limits of the search space.

Then the solution searching process is performed in an organized manner using the following equation

$$u_i = n_{best} + \emptyset_i * (n_{best} - n_i) \quad (3)$$

Where n_{best} is the previous global best solution and $\emptyset_i$ is a random number in the range of [-1,1]. For the first iteration, the first solution is set as n_{best} and the successive iterations take the previous best solution. Thus each penguin generates new solutions and shares the same with its group. The use of the global best solution improves the search operation with maximum exploitation.

After the solutions are determined using the solution search equation, the penguins search and find the local best solutions and update their locations based on the PeSOA update Eq. (1). Then the fitness function is computed using the minimum error of the classifier

Authors Pre-Proof

$$f_j = \frac{f - f_{min}}{f_{max} - f_{min}} \quad (4)$$

Where f_j is the fitness value of j-th feature, f_{min} is the minimum error function and f_{max} is the maximum error function of the classifier. The threshold value for error function is fixed as 0.57. The probability of the selecting a fitness value of j-th feature can be computed by

$$P_j = \frac{f_j}{\sum_{j=1}^{N} f_j} \quad (5)$$

Based on this probability, the features are selected for comparison. The comparison results in the shuffling of the groups of features except the group with minimum error. Then the global best solutions are needed to be computed and hence the information gain is computed for each group to reduce the features. It is computed using the equation

$$Gain(i,j) = entropy\ (i) - entropy(i,j) \quad (6)$$

where $entropy(i)$ is the individual entropy and $entropy(i,j)$ is the average entropy. Entropy can be computed as

$$entropy\ (i) = \sum_{i=1}^{N} -p_i \log_2 p_i \quad (7)$$

Where p_i is the partition class. Finally, the feature groups are ranked based on the information gain values and the best solution is found and update as global best using

$$X_{1\,NEW} = Group\ value\ of\ feature + rand \times (X_{best} - n_{best}) \quad (8)$$

where $X_{1\,NEW}$ is the global best solution, n_{best} the previous iteration best solution and X_{best} is information gain rank value. The pseudo code of improved PeSOA is given as follows:

Pseudo code of the Improved PeSOA:

Read the pre-processed tweet data
Generate random population of P solutions (penguins) in groups;
Initial population of solution n_i can be found using the Eq. (2)
Compute the objective functions for each feature;
Calculate the information gain for each feature (penguin)
Rank the features according to information gain value.
Group the features;
For i= 1 number of generations;
For each individual i ∈ P do
While oxygen reserves are not depleted (stop until 0.00001)
Solution search equation using Eq. (3);
Update the penguin positions using Eq. (1);
Objective function is computed for each group using Eq. (4);
Except the group with minimum error all other groups are shuffled;
Information gain is calculated for each group using Eq. (6).
Rank the features based on information gain value.
Selection of best solution using Eq. (8)
End while
End for
Repeat until best solution obtained.
End

3.5. Classification

The classification of the proposed approach utilizes three classifiers namely k-NN, NB and SVM [43]. The classification performance of these classifiers is improved using the PeSOA and Improved PeSOA. A small description about the classifiers is given below:

Authors Pre-Proof

3.5.1. K-NN classifier

K-NN is the simplest supervised learning classification algorithm, generally used to perform classification and regression processes [43]. It is a neighbor-based lazy classification method that retains training data instances without constructing a model framework for classification. The advantages of this algorithm are its simplicity of implementation, robustness to noisy training data, and efficacy with huge datasets. Nonetheless, k-NN requires the specification of the K value, and the computational expense is significant when the training samples are extensive.

3.5.2. NB classifier

NB classifier is a simple probabilistic classifier based on Bayes hypothesis and is utilized to classify mostly high dimensional inputs [43]. NB classifiers perform effectively in various practical applications, including document categorization and spam detection. The benefit of the NB classifier is the requirement for minimal training data to estimate the essential parameters. NB classifiers are significantly more rapid than more complex methodologies. Nonetheless, they are recognized as poor estimators, rendering them ineffective for estimating tasks.

3.5.3. SVM classifier

SVM represents training data as points in a spatial configuration, categorized by a distinct margin that is maximized. SVMs endeavor to identify the optimal hyperplane that distinguishes positive from negative training samples. The primary advantage of SVM is its efficacy in high-dimensional spaces and its utilization of a subset of training points in the decision function, which enhances memory efficiency. Nonetheless, SVM does not directly yield probability estimates; these are computed through a resource-intensive five-fold cross-validation process.

4. PERFORMANCE EVALUATION

The efficiency of the Improved PeSOA and PeSOA classifiers is compared. The utilized performance metrics are accuracy, precision, recall, F-measure, and processing time. The efficacy of the suggested models is evaluated across two datasets, cancer and pharmaceuticals, with differing data volumes. The cancer tweets are assessed in increments of thousands, ranging from 1000 to 5000, whereas the drug tweets are reviewed in increments of hundreds.

Accuracy: Accuracy is the measure of correctly labeled sentiments in all instances. It can be calculated by

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)} \tag{9}$$

Table 1 presents a comparative analysis of the accuracy between classifiers utilizing PeSOA feature selection and those employing the enhanced PeSOA feature selection. The cancer dataset contains between 1,000 and 5,000 tweets, and the medication dataset has between 100 and 500 tweets. The accuracy of IPeSOA-SVM, after analyzing 5000 tweets in the cancer dataset, is 82.5%, surpassing that of the other methodologies evaluated. Likewise, across all data ranges in cancer and the majority of data ranges in the medication dataset, the IPeSOA-SVM demonstrates superior accuracy. Similarly, the comparison of PeSOA and IPeSOA classifiers indicates that the IPeSOA classifiers exhibit superior accuracy compared to their PeSOA counterparts.

Table.1. Accuracy (%) comparison

Methods	Cancer					Drugs				
	1000	2000	3000	4000	5000	100	200	300	400	500
PeSOA-kNN	77.4	77.8	78.6	78.2	78.9	79.3	79.2	79.2	79.8	78.7
PeSOA-NB	77.3	77.4	78.7	78.2	79.2	79.5	79.3	79.3	79.2	79.0
PeSOA-SVM	79.3	79.6	79.5	79.1	78.7	79.8	79.5	79.5	79.4	79.2
IPeSOA-kNN	79.4	79.2	79.3	79.0	79.1	80.9	80.7	80.7	80.5	80.5
IPeSOA-NB	81.4	81.2	81.7	80.8	80.7	81.2	**83.0**	81.1	**82.0**	81.1

Authors Pre-Proof

IPeSOA-SVM	**82.8**	**82.9**	**83.2**	**82.9**	**82.5**	**82.4**	82.3	**82.2**	81.9	**81.9**

Precision: The precision value is assessed based on true positive predictions and false positives. The calculation of precision is given by

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)} \tag{10}$$

Table 2 presents a comparison of the precision of classifiers based on PeSOA feature selection and those utilizing the enhanced PeSOA feature selection. The cancer dataset contains between 1,000 and 5,000 tweets, and the medication dataset has between 100 and 500 tweets. Likewise, for the majority of data ranges in the cancer and medication dataset, the IPeSOA-SVM demonstrates superior precision values. Furthermore, in the comparison between PeSOA and IPeSOA classifiers, the IPeSOA classifiers exhibit superior precision values compared to their PeSOA counterparts.

Table.2. Precision (%) comparison

Methods	**Cancer**					**Drugs**				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
PeSOA-kNN	78.2	78.8	78.9	79.3	79.2	78.8	79.7	[illegible]	78.7	78.5
PeSOA-NB	78.5	78.2	79.2	79.7	78.7	79.2	79.2	79.2	79.2	79.1
PeSOA-SVM	78.7	78.4	79.4	79.1	79.1	[illegible]	[illegible]	79.4	79.4	79.0
IPeSOA-kNN	80.7	80.5	80.5	80.2	[illegible]	80.5	80.5	80.5	80.5	80.1
IPeSOA-NB	81.1	81.0	81.1	[illegible]	[illegible]	[illegible]	**82.1**	81.1	81.1	**82.0**
IPeSOA-SVM	**82.2**	**81.9**	**81.9**	81.6	**81.6**	**81.9**	81.7	**81.7**	**81.7**	81.2

Recall: The recall value is assessed based on actual positive predictions and false negatives, and is calculated as follows:

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)} \tag{11}$$

Table 3 presents a comparison of recall between classifiers based on PeSOA feature selection and those utilizing the enhanced PeSOA feature selection. The cancer dataset contains between 1,000 and 5,000 tweets, and the medication dataset has between 100 and 500 tweets. In the analysis of 5000 tweets inside the cancer dataset, the recall of IPeSOA-SVM is 71.6%, surpassing that of the other comparative methods. Likewise, for the majority of data ranges in cancer and the medication dataset, the IPeSOA-SVM exhibits superior recall values. The comparison of PeSOA and IPeSOA classifiers indicates that IPeSOA classifiers exhibit superior recall values compared to their PeSOA counterparts.

Table.3. Recall (%) comparison

Methods	**Cancer**					**Drugs**				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
PeSOA-kNN	70.2	68.8	68.7	69.3	69.2	77.7	77.9	76.5	78.2	78.1
PeSOA-NB	68.5	69.0	68.8	68.7	68.7	78.2	78.2	76.7	78.7	78.4
PeSOA-SVM	68.7	68.2	69.2	69.1	69.1	78.7	78.9	77.1	79.1	79.1
IPeSOA-kNN	70.7	70.5	70.5	70.2	70.2	79.1	79.1	79.2	79.2	79.3

Authors Pre-Proof

IPeSOA-NB	71.1	**72.0**	71.1	**71.8**	70.8	80.7	**81.7**	80.8	80.5	80.4
IPeSOA-SVM	**72.2**	71.9	**71.9**	71.6	**71.6**	**81.5**	81.5	**81.23**	**81.35**	**81.2**

F-measure: The F-measure evaluates the accuracy of opinion mining tests and is defined as the weighted harmonic mean of precision and recall. It is provided by

$$F - measure = 2.\frac{precision.recall}{precision+recall} \quad (12)$$

Table 4 presents a comparison of the F-measure between classifiers utilizing PeSOA feature selection and those employing the enhanced PeSOA feature selection. In the analysis of 4000 tweets inside the cancer dataset, the F-measure of IPeSOA-SVM is 83.9%, surpassing that of the other comparative methodologies. In the majority of data ranges within the cancer and medication dataset, the IPeSOA-SVM exhibits superior F-measure performance. The IPeSOA classifiers have superior performance compared to their PeSOA counterparts, as evidenced by elevated F-measure values.

Table.4. F-measure (%) comparison

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
PeSOA-kNN	79.7	79.6	80.8	79.4	81.4	82.2	81.3	81.4	81.3	81.2
PeSOA-NB	79.5	81.1	81.2	81.1	81.2	[illegible]	[illegible]	82.4	82.4	82.3
PeSOA-SVM	82.6	82.2	82.3	81.8	[illegible]	82.8	82.8	82.6	82.5	82.6
IPeSOA-kNN	83.8	83.8	83.6	[illegible]	[illegible]	[illegible]	84.3	84.3	84.45	84.2
IPeSOA-NB	84.3	84.1	84.1	83.8	**83.7**	85.6	85.65	**87.45**	85.5	85.5
IPeSOA-SVM	**84.7**	**84.8**	**84.8**	**83.9**	83.6	**87.41**	**87.3**	87.2	**87.2**	**87.1**

Processing time: It is the complete time taken by the proposed algorithm to provide opinion mining results. The time for processing varies with the size of data evaluated and hence the time for large size tweet files increases.

Table 5 presents a comparison of processing times (in seconds) between classifiers based on PeSOA feature selection and those utilizing the enhanced PeSOA feature selection. The processing time of IPeSOA-SVM for 5000 tweets in the cancer dataset is 18.28 seconds, which is shorter than that of the other methods examined. IPeSOA-SVM exhibits reduced processing time across varying data sizes. It is noteworthy that the IPeSOA classifiers outperform their corresponding PeSOA classifiers.

.Table.5. Processing time (seconds) comparison

Methods	Cancer					Drugs				
	1000	**2000**	**3000**	**4000**	**5000**	**100**	**200**	**300**	**400**	**500**
PeSOA-kNN	10.449	14.51	17.91	20.71	24.19	1.12	2.91	3.53	5.02	5.79
PeSOA-NB	10.332	14.14	17.39	20.39	24.13	0.99	2.81	3.29	4.68	5.58
PeSOA-SVM	10.121	13.97	17.31	20.12	23.92	0.77	2.59	3.32	4.65	5.42
IPeSOA-kNN	5.7351	8.76	12.54	15.45	18.99	1.01	2.28	3.11	4.36	5.05
IPeSOA-NB	5.543	8.42	12.21	**14.88**	18.75	0.91	2.05	3.02	4.14	4.97
IPeSOA-	**5.210**	**8.1**	**11.98**	15.0	**18.28**	**0.76**	**1.98**	**2.87**	**3.99**	**4.66**

Authors Pre-Proof

SVM										

The comparison results indicate that the proposed opinion mining framework, utilizing Improved PeSOA feature selection and SVM classification, demonstrates superior performance, evidenced by elevated accuracy, precision, recall, and F-measure, alongside reduced processing time. The Improved PeSOA algorithm is demonstrably superior to the PeSOA optimization algorithm in the context of opinion mining applications.

5. CONCLUSION

Opinion mining on Twitter is presented in a reasonable and efficient way to interpret timely public sentiment, which is important for decision making in several domains. This research proposed efficient feature selection algorithms for improving the opinion mining performance. The PeSOA is an optimization algorithm inspired by the foraging behavior of penguins, which has been enhanced in this study by modifications in the solution search process and feature reduction utilizing the information gain metric. The classification utilizes three classifiers, and the testing results shown that the enhanced PeSOA significantly improved the classifiers' performance. In the future, the convergence rate of the enhanced PeSOA will be further analyzed to optimize the application of exploitation and exploration properties. The suggested model will also be assessed in other domains to determine its applicability for different uses.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests

REFERENCES

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.
2. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21.
3. Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
4. Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102-107.
5. Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.
6. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.
7. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14-46.
8. Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval*, *12*(5), 526-558.
9. Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*(pp. 1-5). IEEE.
10. Osimo, D., & Mureddu, F. (2012). Research challenge on opinion mining and sentiment analysis. *Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment*, *508*.
11. Rahmath, H. (2014). Opinion mining and sentiment analysis-challenges and applications. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *3*(5).

Authors Pre-Proof

12. Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, *64*(1), 17-27.
13. Carchiolo, V., Longheu, A., & Malgeri, M. (2015). Using twitter data and sentiment analysis to study diseases dynamics. In *International Conference on Information Technology in Bio-and Medical Informatics* (pp. 16-24). Springer, Cham.
14. Zucco, C., Liang, H., Di Fatta, G., & Cannataro, M. (2018). Explainable Sentiment Analysis with Applications in Medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1740-1747). IEEE.
15. Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2018). Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, *68*, 765-777.
16. Ghiassi, M., & Lee, S. (2018). A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*, *106*, 197-216.
17. Saleena, N. (2018). An Ensemble Classification System for Twitter Sentiment Analysis. *Procedia computer science*, *132*, 937-946.
18. Na, J. C., Kyaing, W. Y. M., Khoo, C. S., Foo, S., Chang, Y. K., & Theng, Y. L. (2012). Sentiment classification of drug reviews using a rule-based linguistic approach. In *International conference on asian digital libraries* (pp. 189-198). Springer, Berlin, Heidelberg.
19. Sharif, H., Zaffar, F., Abbasi, A., & Zimbra, D. (2014). Detecting adverse drug reactions using a sentiment classification framework.
20. Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., & Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of biomedical informatics*, *62*, 148-158.
21. Luna-Aveiga, H., Medina-Moreira, J., Lagos-Ortiz, K., Apolinario, O., Paredes-Valverde, M. A., del Pilar Salas-Zárate, M., & Valencia-García, R. (2017). Sentiment polarity detection in social networks: an approach for asthma disease management. In *International Conference on Computer Science, Applied Mathematics and Applications* (pp. 141-152). Springer, Cham.
22. Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *International journal of medical informatics*, *85*(1), 80-95.
23. Crannell, W. C., Clark, E., Jones, C., James, T. A., & Moore, J. (2016). A pattern-matched Twitter analysis of US cancer-patient sentiments. *Journal of surgical research*, *206*(2), 536-542.
24. Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., & Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, *2017*.
25. Kalaivani, P., & Shunmuganathan, K. L. (2015). Feature reduction based on genetic algorithm and hybrid model for opinion mining. *Scientific Programming*, *2015*, 12.
26. Keshavarz, H., & Abadeh, M. S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems*, *122*, 1-16.
27. Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.
28. Iqbal, F., Maqbool, J., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). A Hybrid Framework for Sentiment Analysis using Genetic Algorithm based Feature Reduction. *IEEE Access*.
29. Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, *53*, 453-462.
30. Gupta, D. K., Reddy, K. S., & Ekbal, A. (2015). PSO-ASENT: Feature selection using particle swarm optimization for aspect based sentiment analysis. In *International conference on applications of natural language to information systems*(pp. 220-233). Springer, Cham.
31. Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, *125*, 116-135.
32. Akhtar, M. S., Kohail, S., Kumar, A., Ekbal, A., & Biemann, C. (2017). Feature selection using multi-objective optimization for aspect based sentiment analysis. In *International Conference on Applications of Natural Language to Information Systems* (pp. 15-27). Springer, Cham.
33. Nagarajan, S. M., & Gandhi, U. D. (2018). Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications*, 1-9.
34. Kumar, A., & Khorwal, R. (2017). Firefly algorithm for feature selection in sentiment analysis. In *Computational Intelligence in Data Mining* (pp. 693-703). Springer, Singapore.

Authors Pre-Proof

35. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764-779.
36. Alarifi, A., Tolba, A., Al-Makhadmeh, Z., & Said, W. (2018). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. *The Journal of Supercomputing*, 1-16.
37. Tubishat, M., Abushariah, M. A., Idris, N., & Aljarah, I. (2018). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Applied Intelligence*, 1-20.
38. Kumar, A., & Jaiswal, A. (2019). Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. *Multimedia Tools and Applications*, 1-25.
39. Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of biomedical semantics*, *8*(1), 9.
40. Wegrzyn-Wolska, K., Bougueroua, L., & Dziczkowski, G. (2011). Social media analysis for e-health and medical purposes. In *2011 International Conference on Computational Aspects of Social Networks (CASoN)* (pp. 278-283). IEEE.
41. Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92-111.
42. Gheraibia, Y., & Moussaoui, A. (2013). Penguins search optimization algorithm (PeSOA). In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 222-231). Springer, Berlin, Heidelberg.
43. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

Authors Pre-Proof