M-CAT FEATURE EXTRACTION AND SVM CLASSIFICATION BASED OPINION MINING

[1] **Malathi. M,** [2]**Dr. Antony Selvadoss Thanamani**

[1]*Research Scholar, Department of Computer Science, NGM College, Pollachi, Coimbatore, TamilNadu, India*

[2,] *Head & Associate Professor, Department of Computer Science, NGM College, Pollachi, Coimbatore, TamilNadu, India*

Abstract

Online social media is now an online discourse in which people participate at an impressive rate to build content, post it, bookmark it, and network. Opinion Mining (OM), often referred to as sentiment analysis, is the area of research that analyzes the thoughts, assessments, behaviors, and emotions of people regarding various individuals conveyed in textual input. This is done by categorizing a viewpoint into categories, such as positive, negative, or neutral. In e-commerce websites, opinion mining is very important, and also beneficial to individuals. As a result of user feedback an ever growing amount of results are stored on the web as well as the number of individuals who purchase products from the web increases. Reviews of shipper locations share their thoughts. For example, every organization, web forums, discourse groups, blogs, etc., would have a comprehensive knowledge add-up. Records those are functional for both suppliers and consumers of products on the Site. The method of seeking user opinion on the subject or product or issue is referred to as mining opinion. It can also be described as the process of automatic information extraction, which is called opinion mining, by means of opinions expressed by the consumer who is currently using the software with regard to a certain product. The analysis of the emotions from the opinions extracted is defined as Sentiment Analysis. The goal of opinion mining and Sentiment Analysis is to make computers capable of understanding and expressing feelings. This work concentrates on mining reviews from the websites like Amazon, which allows user to freely write the view. To do this, in this research work Cat Swarm Optimization algorithm, mRMR and Support Vector Machine algorithms are used. To improve the performance of CAT algorithm, it is modified by combining mRMR (Minimum Redundancy and Maximum Relevance) and proposed a new algorithm called M-CAT. Opinion mining is a three step process. In first step pre-processing work is done by Word Stemming, Spelling Check, Letter Replacement, Dialect Replacement. In Second step M-CAT algorithm is used for feature extraction. In third step by using SVM algorithm the result is classified as positive, negative and neutral. At the end we have used quality metric parameters to measure the performance of M-CAT algorithm compared with CAT algorithm.

Keywords*: Modified Cat Swarm Optimization, Opinion Mining, Emotions, Minimum Redundancy and Maximum Relevance, Sentiment Analysis.*

1. INTRODUCTION

Sentiment analysis has become a field of study that is very important. It uses a combination of processes to classify, acquire and distinguish feelings, evaluations, and opinions about persons, subjects, concepts, experiences, information, events, and their characteristics [3]. Sentiment analysis was used to define emerging patterns, interpret user intents and gain understanding. It could be used on multiple sources of data [7]. In many application areas, such as consumer loyalty, political views, predictive analytics and much more, it is very profitable to discover the thoughts and perceptions of users. Sentiment analysis is a natural language processing problem (also referred to as opinion analysis, opinion mining or sentiment mining) [4]. It is a multidisciplinary discipline of study whose theoretical foundations include computer science, linguistics, semantics, and many others. Optimization is the mechanism by which, among many alternative solutions, the optimal solution is chosen for a given problem. The vastness of the search space for many real-life problems, in which it is not feasible to verify all solutions in a reasonable time, is a key issue of this method. Cat Swarm Optimization (CSO) is an algorithm for Swarm Intelligence that was originally invented in 2006 by Chu et al. It is inspired by cats' natural actions and has a new methodology in the simulation of phases of discovery and exploitation [2]. It has been successfully applied in various research and engineering fields of optimization.

2. LITERATURE REVIEW

Computational intelligence is a hot research topic and many related algorithms have been proposed in recent years. Optimization problems are very important in many fields. To the present, many optimization algorithms based on computational intelligence have been proposed, such as the Genetic Algorithm, Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO). In [10], Shu-Chuan Chu, Pei-Wei Tsai proposed a new optimization algorithm, namely, Cat Swarm Optimization (CSO) is proposed. CSO is generated by observing the behavior of cats, and composed of two sub-models by simulating the behavior of cats. According to the experiments, the results reveal that CSO is superior to PSO.

Sentiment analysis is a very substantial area of research. Numerous studies have examined the subject in recent years. It has rapidly gained interest by reason of the unusual volume of opinionated data on the Internet. Significant research has been accomplished to study sentiment by resorting to diverse machine learning techniques. Nevertheless, the downward trend of the accuracy rates in these studies often impacts the approach's efficiency. With the aim of surmounting this obstacle, in [6] Khalid Ait Hadi, Rafik Lasri et al., introduced an efficient technique for sentiment mining in big data context. The data collected are cleaned using a preprocessing data mining technique before proceeding to the selection of the optimal features with the use of a versatile approach of greedy algorithms, called Carousel greedy, combined with a bio-inspired metaheuristic algorithm. The classification is subsequently performed by Cat Swarm Optimization Based Functional Link Artificial Neural Networks classifier and the performance of the approach is discussed through experimental results.

In [2] Aram M. Ahmed, Tarik A. Rashid et al., presents an in-depth survey and performance evaluation of the Cat Swarm Optimization (CSO) Algorithm. CSO is a robust and powerful metaheuristic swarm-based optimization approach that has received very positive feedback since its emergence. It has been tackling many optimization problems and many variants of it have been introduced. However, the literature lacks a detailed survey or a performance evaluation in this regard. Therefore, this paper is an attempt to review all these works, including its developments and applications, and group them accordingly. In addition, CSO is tested on 23 classical benchmark functions and 10 modern benchmark functions (CEC 2019). The results are then compared against three novel and powerful optimization algorithms, namely Dragonfly algorithm (DA), Butterfly optimization algorithm (BOA) and Fitness Dependent Optimizer (FDO). These algorithms are then ranked according to Friedman test and the results show that CSO ranks first on the whole. Finally, statistical approaches are employed to further confirm the outperformance of CSO algorithm.

Today, World Wide Web has brought us enormous quantity of on-line information. In [8] Rasmita Rautray, Rakesh Chandra Balabantaray proposed CSO based model is also compared with two other nature inspired based summarizer such as Harmony Search (HS) based summarizer and Particle Swarm Optimization (PSO) based summarizer. With respect to the benchmark Document Understanding Conference (DUC) datasets, the performance of all algorithms are compared in terms of different evaluation metrics such as ROUGE score, F score, sensitivity, positive predicate value, summary accuracy, inter sentence similarity and readability metric to validate non-redundancy, cohesiveness and readability of the summary respectively. The experimental analysis clearly reveals that the proposed approach outperforms the other summarizers included in the study.

Recently, applications of Internet of Things create enormous volumes of data, which are available for classification and prediction. Classification of big data needs an effective and efficient metaheuristic search algorithm to find the optimal feature subset. Cat swarm optimization (CSO) is a novel metaheuristic for evolutionary optimization algorithms based on swarm intelligence. CSO imitates the behavior of cats through two submodes: seeking and tracing. Previous studies have indicated that CSO algorithms outperform other well-known metaheuristics, such as genetic algorithms and particle swarm optimization. In [7] Kuan-Cheng Lin, Yi-Hung Huang et al., presents a modified version of cat swarm optimization (MCSO), capable of improving search efficiency within the problem space. The basic CSO algorithm was integrated with a local search procedure as well as the feature selection and parameter optimization of support vector machines (SVMs). Experiment results demonstrate the superiority of MCSO in classification accuracy using subsets with fewer features for given UCI datasets, compared to the original CSO algorithm. Moreover, experiment results show the fittest CSO parameters and MCSO take less training time to obtain results of higher accuracy than original CSO.Therefore, MCSO is suitable for real-world applications.

In [4] Harshit Sanwal, Sanjana Kukreja presented opinion mining and summarization of hotel reviews on the web. For opinion classification of hotel reviews we used SVM with Particle swarm optimization (PSO) algorithms Intentions are expressed in a different way with different vocabulary, short forms, and jargon making the data massive and disorganized. The proposed approach is termed sentiment polarity that automatically prepares a sentiment dataset for training and testing to extract unbiased opinions of hotel services from reviews. A comparative analysis was established with compliment Naïve Bayes and Composite hyper cubes on iterated random projections to discover a suitable SVM with Particle swarm optimization(PSO) for the classification component of the proposed approach.

Features are an important source for the classification task as more the features are optimized, the more accurate are results. Therefore, in [3] Dipti Sharma, Munish Sabharwal proposed a hybrid feature selection which is a combination of Particle swarm optimization (PSO) and cuckoo search. Due to the subjective nature of social media reviews, hybrid feature selection technique outperforms the traditional technique. The performance factors like f-measure, recall, precision, and accuracy tested on twitter dataset using Support Vector Machine (SVM) classifier and compared with convolution neural network. Experimental results of this paper on the basis of different parameters show that the proposed work outperforms the existing work.

Nowadays, online social media is online discourse where people contribute to create content, share it, bookmark it, and network at an impressive rate. In [1] Abd. Samad Hasan Basaria, Burairah Hussina et al., attempts to use the messages of twitter to review a movie by using opinion mining or sentiment analysis. Opinion mining refers to the application of natural language processing, computational linguistics, and text mining to identify or classify whether the movie is good or not based on message opinion. Support Vector Machine (SVM) is supervised learning methods that analyze data and recognize the patterns that are used for classification. This research concerns on binary classification which is classified into two classes. Those classes are positive and negative. The positive class shows good message opinion; otherwise the negative class shows the bad message opinion of certain movies. This justification is based on the accuracy level of SVM with the validation process uses 10-Fold cross validation and confusion matrix. The hybrid Partical Swarm Optimization (PSO) is used to improve the election of best parameter in order to solve the dual optimization problem. The result shows the improvement of accuracy level from 71.87% to 77%.

In [9] S. K. Lakshmanaprabu, K. Shankaret al., dissect the high-recommendation web-based business sites with the help of a collection strategy and a swarm-based improvement system. At first, the client surveys of the items from web-based business locales with a few features were gathered and, afterward, a fuzzy c-means (FCM) grouping strategy to group the features for a less demanding procedure was utilized. Also, the novelty of this work—the Dragonfly Algorithm (DA)—recognizes ideal features of the items in sites, and an advanced ideal feature-based positioning procedure will be directed to discover, at long last, which web-based business webpage is best and easy to understand. From the execution, the outcomes demonstrate the greatest exactness rate, that is, 94.56% compared with existing methods. The customer ratings and reviews is very important to the service providers. In [5] K. Sowmya, K. Monika et al., used logistic regression, Naive Bayes, SVM algorithms. We applied these algorithms on the data set containing of 1500 reviews and ratings of the customer. When we see above three algorithms logistic regression is giving 80.82% accuracy, Naive Bayes is giving 67.6% accuracy, where as SVM is giving 80.80% accuracy. When we compare the above classification algorithms accuracy logistic regression and SVM are having good accuracy and better performance.

3. PROPOSED METHOD

Opinion is the opinion of a person reflecting in a specific sense their opinions, beliefs or conclusions in relation to a matter of interest and is generally considered to be subjective in nature. Studies indicate that stakeholder views have a huge effect on decision-making by individuals as well as groups such as governments and organizations rather than evidence. Opinion mining and sentiment analysis, the words that are used interchangeably these days are a field of text data mining that involves extracting opinions from evaluative texts and classifying the polarity of the opinion as positive or negative based on the orientation of the text results after the computational treatment of opinions expressed towards the main features. Natural Language Processing(NLP) techniques are often used in conjunction with KDD methods for different stages of opinion mining, such as opinion statement detection, feature recognition, opinion extraction, polarity determination and opinion summary, since opinions are expressed in human language. Supervised machine learning techniques centered on algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbor (KNN) and Maximum Entropy, among the lexicon-based approaches and machine learning approaches, are widely used to assess polarity for the purpose of classification, using a large number of labeled training data [1]. In this research work mRMR algorithm is combined with Cat optimization algorithm and proposed new algorithm called M-Cat optimization algorithm. Opinion mining process is a three step process. First step is Pre-processing where cleaning process takes place. Second step is feature extraction step, which is very important step for better result. Third step is classification where the result will be opinion like negative, neutral and positive. In this research work SVM algorithm is used for classification.

4. DATA PRE-PROCESSING

Pre-processing is one of the most significant activities in the study of emotion. By reducing its complexity, it can clean the dataset in order to prepare the data for classification. "First, the dataset was tokenized to divide the words into tokens, then stemming will decrease the tokens into a single type, such as decreasing the word "hotels" to "hotel. Redundant words in a text are decreased by the stemming process. Word Stemming - Word stemming involves taking each word into its fundamental form (root). Stemming is the method of reducing a word that adds to suffixes and prefixes or the roots of terms known as a lemma to its word stem. Spelling Check - Social media data contains many spelling errors, such as extra letters. Letter Replacement - Other than English, there are different types for certain letters. The multiple forms of each of these letters have therefore been substituted by some previous research into the default type.

5. FEATURE EXTRACTION

The function is first defined in the collection of features, followed by the selection method and, if necessary, the extraction and reduction process. For identification purposes, feature identification involves recognizing feature forms such as term frequency, term Co-occurrence, Part-of-Speech and Opinion Terms. Several techniques are used for solving this problem of feature subset selection in classification. The major and frequently used approaches are information gain, mutual information, document frequency thresholding, x-test (CHI) and term strength, to list a few. The x statistics and information gain produce good results and are more efficient in optimizing the classification results whereas document frequency is efficient in terms of scalability and complexity. Feature Selection in sentiment analysis is tackling a variety of issues such as large feature space, redundancy, noise attributes, context sensitivity, domain dependency, and limited work on Lexico-structural features, amongst others. The primary objective of the choice of features is to increase the classifier's output by selecting only useful and relevant features and eliminating obsolete, irrelevant and noisy features and thereby reducing the vector function. Further, extracting pertinent and distinct features becomes imperative too when classification algorithms are inept to scale up to the size of feature set in terms of time and space. Absence of proper feature selection technique can cause the classifier to consume more resources and more processing time. The first and foremost challenge in feature extraction is to select the minimal feature subset without any loss of classification accuracy.

A variety of terms as candidate features are considered in a generic emotion classification task, but only a few convey feelings in essence. When they turn the classification process down, these sets of additional features have to be pruned and appear to decrease the classifier's accuracy. Thus, feature selection involves searching optimal feature subset using some search strategies. The search could be exhaustive or approximate, exhaustive search produces optimal solution but it is not feasible for large datasets and the social media data usually have huge dimensionality. Exhaustive search in this case becomes impractical as finding optimal feature subset comes in the category of NP-hard problems as for N number of features, the number of possible solutions will be exponential to 2N. So the focus of researchers has now shifted to meta-heuristic algorithms, which are taken as a subclass of approximate methods. In order to create a more precise classification and minimize the feature set, many evolutionary optimization techniques have been successfully studied and implemented in the past and are also currently being explored, making it a complex research field. Most common evolutionary optimization techniques used for feature selection are, genetic algorithms, simulated annealing, gene expression programming, swarm algorithms, amongst others. In this research work mRMR algorithm is combined with Cat optimization algorithm and proposed new algorithm called M-Cat optimization algorithm. Using this M-Cat algorithm Amazon dataset is preprocessed for better result.

5.1 mRMR

The minimum redundancy and maximum relevance (MRMR) based feature selection algorithm iteratively selects features that are maximally important to the prediction task and minimally redundant to the collection of already selected features, unlike univariate feature selection methods that return a subset of features without accounting for redundancy between the selected features. The following Algorithm 1 shows an mRMR (Minimum Redundancy Maximum Relevance) algorithm.

Algorithm 1: mRMR (Minimum Redundancy Maximum Relevance)

```
CHOSEN ∆ ∅
FS ∆ index list of DS
While FS!= NULL Do
max eval ∆ -1
idx ∆ NULL
For i ∆ 0 To length of FS Do
    tmp flist ∆ {CHOSEN  ∪ FS [i]}
    tmp_eval ⊿ mrmr (tmp_flist)    //evaluate given feature subset by mRMR measure
    IF tmp eval > max eval Then
            max eval ⊿ tmp eval
idx ⊿ i
End If
End For
Append FS [idx] To CHOSEN
Delete FS [idx] From FS
End While
Return CHOSEN
```

5.2 MODIFIED CAT OPTIMIZATION ALGORITHM

A constant and single-objective algorithm is the M-Cat Swarm Optimization. It is inspired by cats' habits of resting and tracing. It seems that cats are lazy and spend much of their time sleeping. However, their knowledge is very strong during their rest and they are very conscious of what is happening around them. So, intelligently and intentionally, they are continually watching the world and when they see a target, they begin to move quickly towards it. The M-Cat Swarm Optimization algorithm is then modeled on the basis of the combination of these two key cats' deportations. The M-Cat Swarm Optimization algorithm consists of two modes: trace and search modes. A solution set that has its own location, a fitness value and a flag is defined by each cat. In the search space, the location consists of M dimensions and each dimension has its own speed; the fitness value shows how good the solution set (cat) is; and finally, the flag is to identify the cats in either the search or tracing mode. Thus, we should first decide how many cats in the iteration should be involved and run them through the algorithm. In each iteration, the best cat is saved in memory and the final iteration cat represents the final solution. The following Algorithm 2 shows the optimization algorithm for M-Cat;.

Algorithm 2: M-Cat Optimization Algotihm

```
Random initialize cats.
WHILE (is terminal condition reached)
Distribute cats to seeking/tracing mode.
FOR ($i$ = 0; $i$ < NumCat; $i$++)
Measure fitness for cat$i$.
IF (cat$i$ in seeking mode) THEN
Search by seeking mode process.
ELSE
Search by tracing mode process.
END
End FOR
End WHILE
Output optimal solution
```

After pre-processing Amazon data, the cleaned dataset is implemented on proposed algorithm called M-Cat algorithm for feature subset selection. In opinion mining selecting feature subset is an important step for better result. M-Cat algorithm based feature selection provides better result for opinion mining.

6. OPINION MINING

Opinion mining and sentiment analysis is a technique used in text documents to identify and extract subjective data. It is a form of text analysis that uses machine linguistics and the processing of natural language to automatically classify and extract feelings or opinions from text (positive, negative, neutral, etc.). Opinion mining is the identification of user opinions from feedback on a specific subject. A relatively new predictor method, both in the case of classification and regression, is the Support Vector Machine (SVM). The Support Vector Machine (SVM) is a collection of directed learning methods used for regression classification and analysis that analyze data and recognize patterns. The original SVM algorithm was created by Vladimir Vapnik. The following Algorithm 3 shows SVM (Support Vector Machine) algorithm. This research work used SVM algorithm for classification after feature set extraction.

Algorithm 3 : Support Vector Machine Algorithm

```
Input: $X$: Training Set    $\delta$: Threshold    Output: $X_R$: $X_R$ C $X$
$|X_R| \ll |X|$  Begin  Train a decision tree $T$; // $X_R$ Begins empty $X_R$ ⊿ $NULL$
For each leaf $L_i$ of T do
    for each opposite class neighbor $L_j$ do
    if entropy of $L_i$ is low then
    //Select closest examples
        Use $L_i$ $L_j$ to build $X^+$;
        Compute $\omega$
        Add $x_j \epsilon L_j$ to $X_R$
        end for
    else
        //Add all the elements in $L_j$ to $X_R$
        $X_R$ ⊿ $X_R \cup L_j$;
    end if    end for return $X_R$ End
```

Support Vector Machine is a rapidly increasing field with promise for greater applicability in all domain of research.

7. RESULT AND ANALYSIS

In this experiment, the Amazon appliances dataset is used. The experiment done on two stages. Initially the dataset is preprocessed using Word Stemming, Spelling Check, Letter Replacement and Dialect Replacement. After preprocessing over feature set is extracted using Cat swarm optimization algorithm. Then this result will be classified using support vector machine algorithm. In second stage the preprocessed dataset is implemented on proposed algorithm called M-Cat algorithm. Then using Support Vector Machine classifier algorithm the result is produced. Finally, the result produced by two algorithms (Cat using SVM and M-Cat using SVM) are compared and proved that M-Cat based SVM opinion mining produced more accuracy, precision, f-measure and recall compared with Cat based SVM opinion mining.

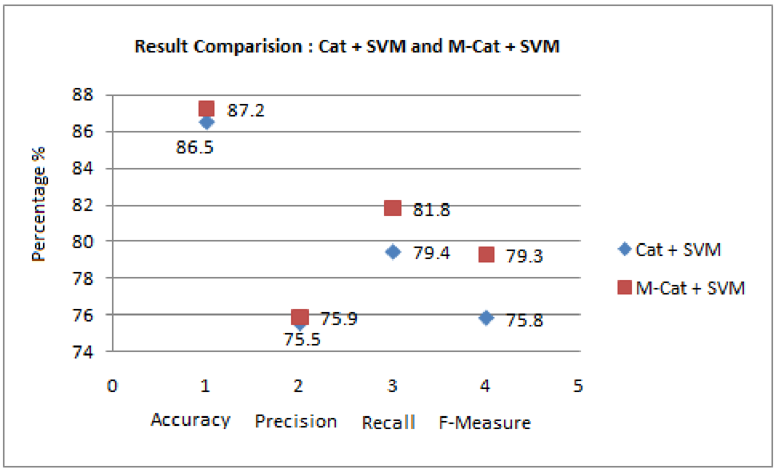

Figure 1: Result Comparisons (Cat + SVM and M-Cat + SVM)

Figure 1. shows the result comparison of Cat + SVM based opinion mining and M-Cat + SVM based opinion mining. The result is based on four parameters such as Accuracy, Precision, Recall, F-Measure. M-Cat based SVM provides 0.7% more accuracy than Cat based SVM. On contrast of precision value M-Cat based SVM provides 0.4% more than Cat based SVM. As well as M-Cat based SVM provides 2.4% more on Recall value. Finally, F-measure of M-Cat scored 3.5% more compared with Cat. It seems proposed algorithm called M-Cat based SVM performs well compared with Cat based SVM on opinion mining.

8. CONCLUSION

Opinion mining and sentimental analysis is an evolving area of data mining used to extract information from a huge volume of comments from consumers, feedback and reviews on any product or topic, etc. In the form of text, sentence and feature level sentiment analysis, a lot of work in opinion mining in customer feedback was conducted to mine opinions. Opinion Mining can be carried out in the future on a collection of discovered feature expressions derived from feedback. Sentiment Analysis is the most interesting research field in Opinion Mining and in the natural language processing community. It is important to invent a more creative and efficient technique to address the current challenges faced by Opinion Mining and Sentiment Analysis. Opinion mining is also known as sentiment analysis. In this paper, according to our experiment, the M-CAT algorithm combined with SVM proves to be the most efficient for text classification of opinion mining.

REFERENCES

1. Abd. Samad Hasan Basaria, Burairah Hussina et al., "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization", Elsevier, Procedia Engineering 53 (2013) 453 – 462, Malaysian Technical Universities Conference on Engineering & Technology 2012, MUCET 2012 Part 4 - Information And Communication Technology.
2. Aram M. Ahmed, Tarik A. Rashid et al., "Cat Swarm Optimization Algorithm - A Survey and Performance Evaluation", Hindawi, Computational Intelligence and Neuroscience, Article ID 4854895.
3. Dipti Sharma, Munish Sabharwal, "Sentiment Analysis for Social Media using SVM Classifier of Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-9S4, July 2019.
4. Harshit Sanwal, Sanjana Kukreja, "Design Approach for Opinion Mining in Hotel Review using SVM With Particle Swarm Optimization (PSO)", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, IJERTV8IS090139, Vol. 8 Issue 09, September-2019.
5. K. Sowmya, K. Monika et al., "Customer Review Rating Analysis Using Opinion Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-7, May, 2019.
6. Khalid Ait Hadi, Rafik Lasri et al., "An Efficient Approach for Sentiment Analysis in a Big Data Environment", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume-8 Issue-4, April 2019.
7. Kuan-Cheng Lin, Yi-Hung Huang et al., "Feature Selection and Parameter Optimization of Support Vector Machines Based on Modified Cat Swarm Optimization", Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, Volume 2015, Article ID 365869, 9 pages.
8. Rasmita Rautray, Rakesh Chandra Balabantaray, "Cat swarm optimization based evolutionary framework for multi document summarization", Elsevier, Physica A 477 (2017) 174–186.
9. S. K. Lakshmanaprabu, K. Shankar et al., "Ranking Analysis for Online Customer Reviews of Products Using Opinion Mining with Clustering", Hindawi, Complexity, Volume 2018, Article ID 3569351, 9 pages.
10. Shu-Chuan Chu, Pei-Wei Tsai, "Computational Intelligence Based on the Behavior of Cats", International Journal of Innovative Computing, Information and Control, Volume 3, Number 1, February 2007, ISSN 1349-4198.