Firefly Algorithm for feature selection in sentiment analysis

Akshi Kumar[1] , Renu Khorwal[2]

Dept. of Computer Science & Engineering
Delhi Technological University
Delhi, India

{[1]akshikumar@dce.ac.in, [2]thekhorwal@gmail.com}

Abstract. Selecting and extracting feature is a vital step in sentiment analysis. The statistical techniques of feature selection like document frequency thresholding produce sub optimal feature subset because of the Non Polynomial(NP) hard character of the problem. Swarm Intelligence algorithms are used extensively in optimization problems. Optimization techniques could be applied to feature selection problem to produce Optimum feature subset. They render feature subset selection by improving the classification accuracy and reducing the computational complexity and feature set size. In this work, we propose Firefly algorithm for feature subset selection optimization. SVM classifier is used for the classification task. Four different datasets are used for classification of which two are in Hindi and two in English. The proposed method is compared with features selection using genetic algorithm. This method therefore, is successful in optimizing the feature set and improving the performance of the system in terms of accuracy.

Keywords: Sentiment Analysis, Feature Selection, Swarm Intelligence, Firefly Algorithm, Genetic Algorithm, Support Vector Machine(SVM)

1 Introduction

With the advent of social media, opinion-rich data resources such as microblogging sites, personal blogs and online review sites have proliferated enormously. People express their views or opinions/attitudes on a variety of issues, discuss current issues, complain, and provide feedback and suggestions for the products and policies they use in their daily life or which concerns them. This unstructured social media data is used to mine the overall attitude of the writer towards a specific issue. Sentiment Analysis or Opinion Mining [1], as an intelligent mining technique, helps to capture & determine opinions, emotions and attitudes from text, speech, and database sources, which correspond to how users retort to a particular issue or event. Being a Natural Language Processing task it tries to gauge the opinion of writer about the issue at hand and examine the overall contextual polarity of the data. Sentiment mining from social media content is a tedious task, because it needs in-depth knowledge of the syntactical & semantic, the explicit & implicit, and the regular & irregular language rules.

Sentiment Analysis is a multi-step process encompassing various sub-tasks, that are, Sentiment Data collection; Feature Selection; Sentiment Classification and Sentiment Polarity detection [2]. Feature selection in sentiment analysis has a very significant role in enhancing accuracy of the system as the opinionated documents usually have high dimensions, which can adversely affect the performance of sentiment analysis classifier. Effective feature selection technique recognizes significant and pertinent attributes and improves the classification accuracy thereby reducing the training time required by classifier. Due to the high- dimensional, un-structured characteristics of the social media content, this problem of text classification manifolds, thus fostering the need to look for improved & optimized techniques for feature selection.

The Traditional methods for feature selection that are chi-square, information gain, and mutual information etc. [3] are successful in reducing the size of the corpus but with a compromised accuracy. These produce sub-optimal feature subsets as feature subset selection problem lies in the category of Non-Polynomial (NP) hard problems. Evolutionary algorithms have been successful at coming up with good solutions for complex problems, when there is a way to measure quality of solutions [7]. Algorithms such as Nature-Inspired Algorithms [6], Genetic Algorithms [5], Simulated Annealing [8], etc. have been explored much in literature for improved classification. Swarm Intelligence is a distributed system whereby self-cooperating global behavior is produced by anonymous social agents interacting locally having local perception of its neighboring agents and the surrounding environment. These algorithms work on the principle of distribution of labor and distributed task allocation producing global patterns, the individual agents such as bees, ants, can do their individual task while it is the cooperative work of whole colony brings out intelligent behavior [4].

Firefly algorithm, developed by Xin-She Yang [14] in the year 2008, is a biologically inspired algorithm centered around flashing patterns of fireflies. Firefly global optimization is a relatively new algorithm and outperforms other swarm intelligence algorithms. FA have attracted a lot of researchers in the recent years. It is a Population based metaheuristic algorithm based on pattern of fireflies where each firefly represents potential solutions to the problem in the search space. The proposed system uses firefly optimization algorithm for feature selection. The proposed technique is compared with feature selection using genetic algorithm and the baseline model and is validated for four different datasets.

2 Related Work

SA has received a lot of focus from researchers, analysts in recent years. T.Sumathi et al[11] (2013) introduced feature selection technique for selecting optimum feature set thereby improving classification accuracy and reducing computational complexity. ABC algorithm is used for feature selection optimization. The ABC technique is a powerful optimization technique and is widely used in optimizing NP Hard problems. Opinion mining is used for classifying movie reviews where features are optimized using ABC algorithm. This method improved classification accuracy in tune of 1.63% to 3.81%. Ruby Dhruve et al[9] investigated Artificial Bee Colony algorithm for calculating weight of sentiment. The proposed technique incorporated ABC algorithm for classification to improve the accuracy of the classifier with BOW and BON features. Sphere benchmark function is used is this work for optimizing the best result of classification. Experimental results show an accuracy improvement from 55% to 70%. Particle Swarm Optimization was used in [13] for feature selection in aspect based sentiment analysis. The proposed technique here could automatically determine the most relevant features for sentiment classification and also this works focus on extraction of the aspect terms. Experiments revealed that the system is able to achieve better accuracy with a feature set having lower dimensionality. A comparative analysis of work done in feature selection using evolutionary algorithms is given in Table 1.

Technique	**Dataset**	**Classifier**	**Accuracy without optimization**	**Accuracy with optimization**	**Year**	
ABC	Product Reviews	SVM	55	70	2015	[9]
ABC	Internet Movie Database (IMDb)	Naïve Bayes	85.25	88.5	2014	[11]
		FURIA	76	78.5		
		RIDOR	92.25	93.75		
hybrid PSO/ACO2	Product Reviews, Government data	Decision Tree	83.66	90.59	2014	[10]
PSO	Twitter Data	SVM	71.87	77	2012	[12]
PSO	Restaurant Review	CRF	77.42	78.48	2015	[13]

Table 1: Analysis of feature selection optimization techniques

Distance based discrete firefly algorithm is used in [15] for optimal feature selection using mutual information criterion for text classification. The system proposed in this work uses mutual information based criterion which can measure association between two features selected by the fireflies and determines corrections of features. Also the system is compared with genetic algorithm, two PSO variants and differential algorithm. The work done in this system produces results having greater accuracy.

3 Methodology

The proposed system selects optimum feature subset from high dimensional feature set for sentiment analysis using firefly algorithm and also the results are compared with genetic algorithm. The fundamental idea of firefly algorithm is that the fireflies uses information about brightness from its neighbors to assess themselves. Each firefly is attracted towards its brighter neighbor based on distance. In the standard firefly algorithm, the search strategy is reliant on its control parameters i.e. absorption parameter and randomness parameter. Optimality degree and the time required to obtain the optimality are two important evaluation aspects of the feature selection problem. The current methods of feature subset selection are successful in achieving either of the two criteria but not both. So, firefly algorithm is used here to tackle both the problems simultaneously. The algorithm used is the binary version of the discrete firefly as it deals with candidate solutions in terms of n-bit binary string values.

3.1 Feature Selection Using Firefly optimization algorithm

A discrete FA is proposed in this section to solve the feature selection problem. Pseudo code for the proposed system is presented in Algorithm1. The fitness function f is used to measure the fitness based on accuracy of particular solution $\chi_{i.}$

```
Feature Selection Using Firefly Algorithm
Input : N number of fireflies
        $T_{max}$ Maximum number of iterations
        $\gamma$ Absorption parameter
        $\alpha$ Randomness Parameter(environment noise)
Output : Optimal firefly position and its fitness
  1. Initialize parameters N, $T_{max}$, $\gamma$ and $\alpha$
  2. Initialize $\chi_{i} = \varphi$ , subset of feature selected by $i^{th}$ firefly
  3. Initialize $x_i$, position of each firefly subjecting to $\Sigma\, x_{ij} = s$
  4. Calculate fitness $f(\chi_i)$
  5. Sorting the fireflies in accordance with $f(\chi_i)$
  6. While $t < T_{max}$
       for i = 1 to N (for each firefly)
          for j = 1 to N (for each firefly)
             if $f(\chi_j) > f(\chi_i)$
                move firefly i towards firefly j using equation
             end if
           update $\gamma$ and corresponding attractiveness
          end for
        end for
       evaluate the position of fireflies
         t=t+1
     end while
  7. Output the feature subset
```

Algorithm 1: Pseudo code of the Firefly algorithm

Step 1. The first step is to initialize the firefly parameters size for the firefly population(N), α- the randomness parameter, γ-absorption coefficient, and t_{max} - maximum number of generations required for termination process.

Step 2. Next initial firefly position is initialized. Initially a random position is allocated to fireflies. $X_i = [x_{i1}; ...; x_{in}]$ where n represents number of features in total and m solutions are considered as candidate solutions initially. Each candidate solution X_i is represented as a binary string vector.

Step 3. In the next step the fitness of the population is calculated. For this case the function of fitness is the accuracy of the classifier model which is the accuracy of the SVM classifier with current solution as feature set.

Step 4. Firefly position modification. The firefly having less brightness value would move towards a firefly with more brightness. The new position is based on modification in each dimension of firefly.

Step 5. The new solution produced by modifying firefly position is examined using the fitness value of the SVM classifier. Solutions with very low accuracy are discarded

Step 6. Store the best solution attained till now and increment the generations counter.

Step 7. If the termination criteria are satisfied, then we stop the search otherwise, go to step 3. The termination criteria used in this method is maximum number of generations or if classification error is negligible.

Initial Population and encoding of fireflies

Initial population is generated randomly as array of binary bits, with length equal to the total number of features. The potential solution has the form $\vec{X}(i) = (x_{i1}, x_{i2}, ..., x_{in})$ where $x_{ij} \in \{0, 1\}$, i = 1, 2, ..., N, where N specifies the number of feature set i.e. the population size and j = 1, 2, ...n where n is the number of features. Firefly length is equal to total number of features. Suppose we have a feature set F = (f1, f2, f3, ..., fn), then a firefly is represented as a binary vector of length n. In $\vec{X}(i)$ if any bit holds a value of "0" then the particular feature is not used for training the classifier, and a value "1" indicates that the corresponding feature is used for classification. We define the number of features we want to use for classification. For e.g., Considering a set of features, F = (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}) and taking N the total population size as 3, the firefly population can be symbolized as follows:

$$\vec{X}(1) = (1, 1, 0, 1, 0, 1, 0, 1, 0, 1)$$
$$\vec{X}(2) = (0, 1, 1, 1, 0, 0, 1, 1, 0, 1)$$
$$\vec{X}(3) = (1, 0, 0, 1, 1, 1, 1, 0, 1, 0)$$

The initial population is generated randomly for N solutions of feature sets. In the initial feature set the bit positions are randomly assigned as 1 or 0. This is done by generating a uniform random number c in range [0,1] for every bit position of the feature vector string i.e., X_{id} of $X(i)$. Based on the random value every firefly $X(i)$ is created as follows:

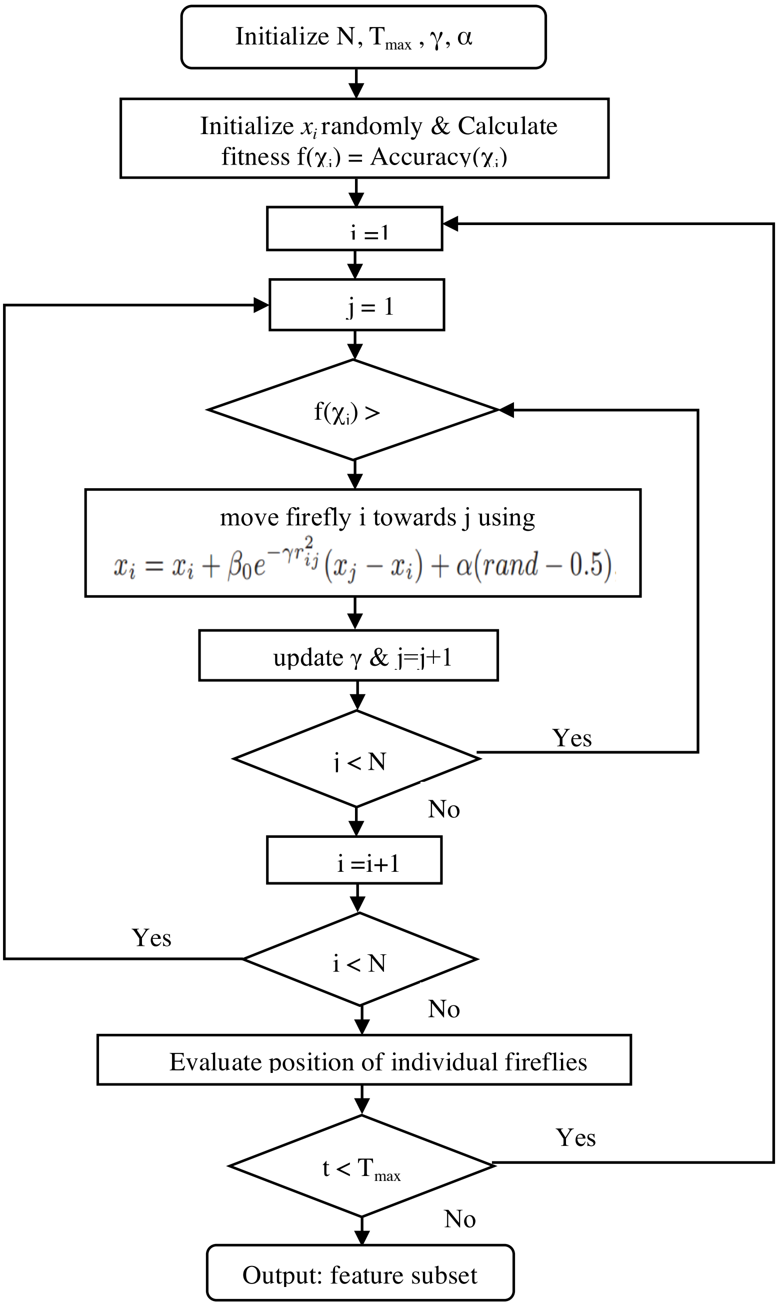
$$X_{id} = \begin{cases} \mathbf{1} & \boldsymbol{if\, c} \textbf{ is } < \mathbf{0.5} \\ \mathbf{0} & \textbf{otherwise} \end{cases} \quad (1)$$

Updating the Global and Best Firefly Position Value

The best particle from the initial population is first selected based on the fitness value i.e. the accuracy of the classifier. Now the firefly having low value moves towards the firefly having more value of the brightness i.e. fitness function(accuracy).

$$p_{ij} = \frac{1}{1+e^{\vartheta_{ij}}} \qquad \textbf{(2)}$$

Initially the local best and global best are set as same which is the best firefly position from the initial population. The terms in equation (3) are for continuous optimization to use it in discrete optimization the terms are required to be converted into discrete form. So they are converted into discrete form using function given in equation (2).



Initialize N, T_{max} , γ, α
Initialize x_i randomly & Calculate fitness f(χi) = Accuracy(χi)
i =1
j = 1
f(χi) >
move firefly i towards j using
$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha(rand - 0.5)$
update γ & j=j+1
j < N
Yes
No
i =i+1
i < N
Yes
No
Evaluate position of individual fireflies
t < T_{max}
Yes
No
Output: feature subset

Fig 1: Flow chart of firefly algorithm for Feature selection

P_{ij} in equation (2) is the probability that j^{th} bit is set in x_i. The movement of a firefly *i* towards firefly *j* based on attraction which is computed using equation (3). Here rand is a random number generated in range [0,1]. The second term in equation is because of the attraction between the two fireflies due to brightness variance and the third term used is for bringing randomization to make exploration.

$$\vartheta_{ij} = \beta_0 \mathrm{e}^{-\gamma r_{ij}^2}(x_{kj} - x_{ij}) + \alpha\left(\mathbf{rand} - \frac{1}{2}\right) \tag{3}$$

The i^{th} firefly uses the update rules that are given by equation (4). In this equation rand is used to denote a random number generated on interval (0,1). Specific value could also be used instead of the rand value used in equation (4).

$$x_{ij}^{t+1} = \begin{cases} 1 & \text{if } p_{ij} \geqslant \text{rand} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Parameters controlling exploration and exploitation

The two parameter randomness and attractiveness controls the quality of solutions produced and also the convergence rate. Selecting an optimum value of these parameters is an open area for research. The α i.e. the randomness parameter affects the light transmission, it can be varied for producing variations in the solution and providing more diversity in candidate solution space. It basically represents the noise existing in the system, $\alpha \in [0, 1]$. The parameter γ also known as absorption coefficient characterizes the variation of the attractiveness, and its value is vitally important for evaluating the convergence speed of algorithm [16]. It controls exploration and exploitation of algorithm and is generally selected in range [0, ∞]. If $\gamma \rightarrow \infty$, the brightness and attractiveness will go down significantly, which bases the fireflies getting lost in the search process. Whereas, if $\gamma \rightarrow 0$, the brightness and the attractiveness will be constant. Proper setting of these variable can enhance the performance of the firefly algorithm significantly.

Feature Selection Using Genetic algorithm

Genetic algorithm starts with an array of population of candidate solutions where, each solution in the initial population is encoded with genes and each gene represents individual feature. The presence of feature is marked by gene value 1 and feature is not present if gene value is 0. New population (feature set) is generated from the previous population and the features with good accuracy are used for producing the new population. Parents are selected from previous population and are mixed using crossover and mutated randomly to produce new and improved offspring feature sets. Algorithm 2 represents the pseudocode for genetic algorithm in feature selection for sentiment analysis.

4 Implementation and Results

4.1 Corpora Description

To conduct this work movie review and Twitter data extracted from twitter API are considered here. Dataset in two languages Hindi and English is considered. The movie review dataset of Hindi language has been taken from Resource Centre for Indian Language Technology Solutions(CFILT), IIT Bombay Sentiment Analysis Group Resources (http://www.cfilt.iitb.ac.in/Sentiment_Analysis_Resources.html).

```
Feature Selection Using Genetic Algorithm
    Input : P Initial population
            $P_{size}$ Population size
            $g_{max}$ Maximum number of generations
    Output : Fittest chromosome bestp
        1. Initialize $t = 0$ & initialize initial population $P(t)$ randomly
        2. Compute fitness $f(t) = accuracy(P(t))$
        3. While $t < g_{max}$ do
               $t = t + 1$
               for individual (i) in P (for each individual)
                   $P_i(t) = crossover(P_i(t - 1))$
                   $P_i(t) = mutate(P_i(t))$
                   $F_i(t) = accuracy(P_i(t))$
               end for
             update bestp
           end while
        4. Output the feature subset bestp
```

Algorithm 2: Pseudo code of the Genetic algorithm

Dataset	**Language**	**Number of Samples**	**Positive reviews**	**Negative reviews**
Movie review	English	1000	500	500
Movie review	Hindi	302	127	125
Twitter data Keyword-“Delhi Government”	English	800	400	400
Twitter data Keyword- “□□□□□□ सरकार”	Hindi	500	250	250

Table 2: Dataset Used

The Firefly parameters adapted are as follows:
1) Firefly population size: 50
2) Firefly length= Total number of features
3) Number of generation: 50
3) Absorption coefficient α: 0.5
4) Attractiveness parameter γ: 0.9

The result of classification of SVM classifier, SVM-Genetic classifier and SVM-firefly classifier for the four datasets used in this work is given below in Table 2. Comparisons with baseline SVM system and optimized SVM systems shows that the optimized system produce more accuracy as compared to the baseline model. This shows that promising accuracies with much reduced feature set can be achieved using evolutionary optimization.

Model	**Accuracy**			
	Movie Review (English)	**Movie Review (Hindi)**	**Twitter Data keyword- {‘Delhi’, ‘government’}**	**Twitter Data keyword- {‘□□□□□□’, ‘सरकार’}**
SVM	79.55	74.12	80.3	73.5
Genetic-SVM	82.15	77.31	83	77
Firefly- SVM	85.29	79.6	86.71	78.46

Table 3: Average percentage improvement in accuracy

The firefly algorithm is a powerful optimization algorithm and produces great improvement in accuracy in our model. The genetic algorithm which is also an evolutionary algorithm model is better than the baseline model but less efficient than the firefly model of feature extraction. The system is validated with four different datasets of which two are in Hindi language and other two in English language.

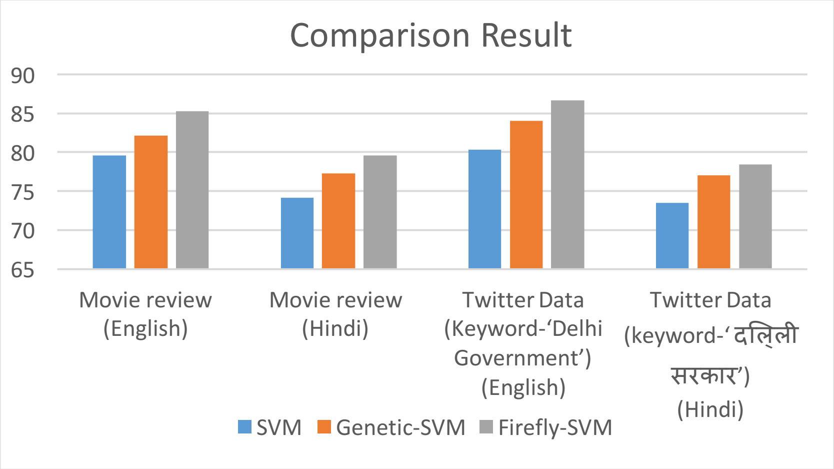


Fig 2: Comparison results of Firefly method, Genetic method and baseline SVM

Conclusion

Sentiment analysis systems are used in almost every domain be it social, business or political as opinions are key influencers of almost all human activities. Formed by perception these opinions guide our conducts in various domains. Sentiment analysis finds many applications few of them being product Perception: evaluating customers' sentiments towards particular product and evaluating trend change over time, identifying feedback over various products and policies to define new targets for marketing. Accuracy of sentiment analysis systems is very important for these applications. Reduction of the feature vector size considerably leads to improvements in accuracy as there are features which are noisy and redundant and not required for classification. In this work feature subset reduction is achieved using firefly algorithm. It improves the accuracy of sentiment analysis considerably by reducing the feature set size. The hybrid firefly-SVM model brings an accuracy improvement of 5.64 on average which is significant improvement. The experiment results reveal that the classification accuracy increases on an average by 3% in case of Genetic algorithm and by 5.64% in case of Firefly algorithm. Also the hybrid method works very well for languages other than English as shown in this work where hybrid method works well for Hindi language also.

References

1. Bo Pang., Lilliam Lee.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–13 (2008)
2. Akshi Kumar, Teeja Mary Sebastian, Sentiment Analysis: A Perspective on its Past, Present and Future, International Journal of Intelligent Systems and Applications, Vol.4, No.10, 2012
3. Yiming Yang, Jan O. Pederson, A Comparative study on Feature Selection in Text Categorization(1997)
4. Mehdi Hosseinzadeh Aghdam *, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, Text feature selection using ant colonyoptimization, Expert Systems with Applications 36 (2009) 6843–6853
5. Ekbal, A., Saha, S., and Garbe, C. S. "Feature selection using multi objective optimization for named entity recognition" In proceedings of IEEE 20th International Conference on Pattern Recognition, pp. 1937-1940,2010.
6. Sangita Roy, Samir Biswas, Sheli Sinha Chaudhuri, Nature-Inspired Swarm Intelligence and Its Applications, I.J. Modern Education and Computer Science, 2014, 12, 55-65
7. Carlos M. Fonseca, Peter J. Fleming. An Overview of Evolutionary Algorithms in Multiobjective Optimization. Spring 1995, Vol. 3, No. 1, Pages 1-16 Massachusetts Institute of Technology, Online December 10, 2007.
8. William L. Goffe, Gary D. Ferrier, John Rogers, Global optimization of statistical functions with simulated annealing, Journal of EconometricsVolume 60, Issues 1–2, 1994, pp 65-99
9. Ruby Dhurve, Megha Seth, " Weighted Sentiment Analysis Using Artificial Bee Colony Algorithm", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064
10. George Stylios, Christos D. Katsis, DimitrisChristodoulakis, " Using Bio-inspired Intelligence for Web Opinion Mining", International Journal of Computer Applications Vol 87 – No.5, 2014
11. T. Sumathi, S.Karthik, M.Marikkannan, "Artificial Bee Colony Optimization for Feature Selection in Opinion Mining", Journal of Theoretical and Applied Information Technology, 2014. vol. 66 no.1
12. Abd. Samad Hasan Basari, BurairahHussin, I. GedePramudya Ananta, Junta Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization"
13. Deepak Kumar Gupta, Kandula Srikanth Reddy, Shweta, Asif Ekbal, "PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis", Natural Language Processing and Information Systems,Volume 9103 pp 220-233

14. Xin-She Yang. Firefly algorithm, stochastic test functions and design optimization. International Journal of Bio-Inspired Computation, 2(2):78–84, 2010.
15. Long Zhang, Linlin Shan, Jianhua Wang, Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion, Neural Computing and Applications, ISSN 1433-3058, Springer, 2016
16. Xin-She Yang, Xingshi He, Firefly Algorithm: Recent Advances and Applications, International Journal of Swarm Intelligence, 2013 Vol.1, No.1, pp.36 – 50