International Journal of Engineering &Technology, 8 (4) (2019) 582-589

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Feature selection using ant lion optimization algorithm in text categorization

B. Sunil Srinivas[1]*, A. Govardhan[2]

[1]Assistant Professor, CSE Dept., TKR College of Engineering & Technology, Hyderabad, & Research Scholar, CSE Dept., JNTUA
[2]Professor, CSE Dept., JNTU College of Engineering, JNTUH, Hyderabad
**Corresponding author E-mail:sunilsrinivas16@gmail.com*

Abstract

This is Big Data decade with extensive increase in the textual information where the text classification is the significant approach for processing and organizing textual information. Text categorization refers to the process of spontaneously allotting documents to the relevant classes. The key features of these text classification issue is tremendous increase in higher dimensionality of text information. Meta-Heuristics Approaches are effortlessly employed to obtain optimal solutions for high dimensional datasets in text categorization. However, some of these approaches like genetic algorithm and particle swarm optimization gives a sub-optimal solutions, the convergence time is more compared to other approaches and cannot guarantee the global maxima to the text categorization. Thus, in this paper, a nature-inspired optimization approach depending on catching mechanism of antlions in the environment known as Ant Lion Optimizer (ALO) Approach, is applied to resolve higher dimensionality issues prior to text classification. The precision and recall values for the proposed is comparatively effective when compared with the existing text categorization dimensionality reduction techniques.

Keywords:*Antlion Optimization Algorithm; Classification; Dimensionality Reduction; Feature Selection Text Categorization; Support Vector Machine*

1. Introduction

The rapid increase in the internet usage and availability of on-line documents has made the job of processing textual information as one of the key issue now a days. With this growing volumes of data in internet and commercial intranets, there is an increasing requirement for devices that assist individuals to filter and organize the database. Text categorization [1] is defined as the job of spontaneously allotting documents to the relevant classes. Till late ’80s, the utmost significant approaches depends on knowledge engineering that is physically determining the set of guidelines encrypting skilled knowledge. Currently, the finest text categorization employs the machine learning algorithm that is the categorizer acquires guidelines from instances, and calculates on the group of trial documents. Nevertheless, with the rapid rise in the web data sample, approaches that can enhance the classification efficacy along with the accuracy are extremely preferred.

A dominant issue in numerical textual categorization [2] is the higher dimensionality in the feature domain. A single dimension is present for every distinct word obtained from the group of documents characteristically hundreds of thousands. Alternatively, certain unnecessary and recurrent features might injure the predicting performance of categorizers for textual classification. Usual classification methods could not work with the huge feature data sets, as handling and organization is extensively expensive in terms of computation where outcomes has been undependable because of the deficiency of adequate training samples. Therefore, there is a requisite for the reduction of original feature sample that is usually called dimensionality reduction approach. The selection of certain illustrative characteristics from original feature domain is required to diminish the size of feature set and to enhance the efficacy and accuracy of categorizers.

The Conventional approaches for selection of feature subset in text categorization employs an estimation function that can be smeared on every unique terms. The complete set of terms are autonomously estimated and arranged as per the allotted strategy. Further, the numerous pre-specified finest characteristics are considered to the built the finest feature subset. Evaluation of each single terms could be handled using certain metrics such as document rate [3], mutual information [3], information gain or $\chi 2$ statistic ([3 - 5]) and odds-ratio [6].

The application of machine learning (ML) and Pattern Recognition (PR) in practice, information frequently come across the issues initiated due to higher magnitude of input domain [7]. In the current era, metaheuristic approaches are used as the key approach for attaining the optimum outcomes to dimensionality reduction problem. The time and space complexity in evaluation is being challenged by numerous arithmetical approaches however inferior than the biologically inspired approaches. Numerous approach are suggested to tackle the issue of dimensionality reduction particularly meta-heuristics approach. Among them, genetic algorithm and particle swarm optimization algorithm are most well-known methods. However, these approaches also suffer from certain limitation.

These approaches sometimes provide sub-optimal solutions, the convergence time is more compared to other approaches and cannot guarantee the global maxima. The closeness of subsequent solutions with each other dependent on the convergence of these approaches. In order to eliminate some of these limitations and select a sub group of accessible features through removing inappropriate features for

Copyright ©B. Sunil Srinivas, A. Govardhan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

the categorization task, the proposed approach suggested novel meta-heuristics based dimensionality reduction approach for text categorization. The antlion optimization algorithm is employed for this purpose. This is a recent meta-heuristic that mathematically models the interaction of ants and ant lions in nature.

1.1. Organization of the chapter

A brief introduction to text categorization along with motivation for the suggested methodology is given in this section. The section 2 briefly gives the survey of existing text categorization approach and numerous dimensionality techniques in text categorization. The Antlion Optimization approach is explained briefly in section 3. The Antlion optimization algorithm based dimensionality reduction for text categorization is briefly deliberated in section 4. The experimental results and its analysis is elucidated in detailed in section 5. In section 6 and section 7, the conclusion and references are given.

2. Literature survey

A review of cross-domain text categorization problem are shown in [8]. Unlike the classical case, the training and the test data originates from diverse distributions or provinces. This is very common in practical tasks because (especially for Polish language) we often do not have a suitable data set of labelled documents. Often what we have is a corpus which is topically related, but presents the same (or semantically similar) information in a different way, e.g. using different vocabulary. Many algorithms have been developed or adapted for cross-domain text classification, there are conventional algorithms

In [9], a general overview of the problem of semantic gap in information retrieval is given. Authors focus on two separate task: text and multimedia mining/image retrieval. Semantic gap in text retrieval is defined as a usage of different words (synonyms, hypernyms, and hyponyms) to describe the same object. In the part about text retrieval authors concentrate on reorganizing search results by using post-retrieval clustering system. They work on search results ("snippets") and enhance them by adding so called topics. Topic is a set of words (they have similar meaning) that was as outcome of Probabilistic-Latent Semantic analysis or Latent Dirichlet Allocation on some external data collection. After adding a topic to the snippet they carry out clustering or labelling.

In [10], authors suggested a methodology to enhance the classification through accumulating semantic information from Wikitology (knowledge repository based on Wikipedia). They used various text representation and text enrichment techniques and used Support Vector Machine-SVM to learn a prototype for categorization. Forman [11] gives a widespread comparative analysis of twelve feature selection strategy for higher dimensional domain. Waqas et al. [12] focused on multi-objective GA for solving feature subset selection. The research showed that independent subsets of features are excellent in accuracy.

AlSukker et al. [13] presented a novel modified genetic approach depending on improved population diversity, parents' choice, and enhanced genetic operations. Real-world outcomes signified the importance of suggested GA variation matching with numerous approaches from survey on diverse data samples. Mahrooghy et al. [14] employed filter based feature selection genetic algorithm (FFSGA) to obtain an optimum group of characteristics where recurrent and inappropriate features are eliminated. The entropy indexed fitness evaluation was employed to estimate feature subsets. The outcome exhibited that employing feature selection method not merely enhanced the justifiable threat score with nearly 7% at certain threshold values in winter, however extensively minimized the size.

3. Ant lion optimization algorithm

The Ant Lion Optimizer, known as ALO or Antlion Optimizer [8], is a recent meta-heuristic that mathematically models the interaction of ants and antlions in nature. Recently, Mirjalili proposed an Antlion Optimizer (ALO) algorithm on the behaviour inspired from antlions. This approach depends on foraging behaviour of antlions. Moreover, salient features of algorithms are the effective exploration of search domain through random walk and random selection of agents. Similarly, exploitation of the search domain is assured through adaptive borders of traps. Since it is a population aided approach, the avoidance of local optima is indispensable.

Antlions refers to the collection of bugs in the group of Myrmelentidae. Two key stages of antlions lifecycle are larval and adult. The larva is frequently known as "doodlebug" due to the traces it leaves in sand whereas exploring for finest position to construct the trap. In course of this approach of hunting, antlion makes funnel holes in lenient sand and further waits unwearyingly towards end of the hole. Sliding to the end of the hole, the prey is instantaneously grabbed by antlion. Otherwise, if prey tries to escape from trap, antlion throw sands in the direction of the end of hole to slip the prey inside the hole. By throwing up soft sand inside the hole, the larva similarly extract the edges of the hole, triggering them to breakdown and fetch the prey towards itself. Whenever a prey is trapped into the jaw, it is dragged inside the soil and consumed. Once, prey is consuming by antlions, it throws the remaining sand into hole and modify hole for subsequent explore. The other fascinating behaviour witnessed in life cycle of antlions is the relevancy of dimension of trap along with the level of starvation and outline of moon. Antlions usually dig huge traps since they are in starvation whenever moon is full. They have been progressed and altered the behaviour to enhance its chance of existence. It is similarly found that an antlion do not straight witness the shape of moon so as to determine the dimension of the trap, however has an interior lunar clock as to make the decisions .The key motivation of ALO approach obtains from foraging behaviour of antlions larvae.

The antlion optimizer performs on the two stages:

i) Constructing a trap: Roulette wheel is employed to prototype the exploring ability of antlions. Ants are presumed to be trapped in merely preferred antlion pit. The ALO approach needs a roulette wheel operation for picking antlions depending on its fitness in course of optimization. This technique gives higher probabilities to fit antlions for grasping the prey.

ii) Grasping prey and re-constructing the pit: This is the last phase in exploration, where the antlion takes the ant. It is presumed that prey grasping happens whenever the ant be fittest compared to its equivalent antlion. The antlion need to update its location to modern location of exploring ant to augment its chance of grasping novel prey. Eq (1) represent this approach:

$$Antlion^{t}_{j} = Ant^{t}_{i} \ if \ f(Ant^{t}_{i}) is \ superior \ than \ f(Antlion^{t}_{j}) \quad (1)$$

Here t exhibited the present generation, $Antlion^{t}_{j}$ exhibited the location of antlion j at the generation t and Ant^{t}_{i} signifies the location of ant i in the generation t.

The antlion optimizer functions using four stages for an individual ant:

iii) Sliding ants in the direction of antlion: Antlions shoot sand in the direction of center of pit once the ant is in trap of antlion. This actions makes the trapped ant to escape from sliding down. To arithmetically prototype this behavior, the radii of ant's arbitrary movement hyper-sphere is minimized adaptively employing Eqs (2) and (3).

$$c^t = \frac{c^t}{I} \quad (2)$$

Here $c^t$is the minimal of entire variables at generation t and I is a ratio that is specified in Eq (3)

$$I = 10^w \frac{t}{T} \quad (3)$$

Here t is the present generation, T is the maximal number of generations, w is a constant determined depending on present generation. Fundamentally, the constant w could alter the accuracy level of utilization.

iv) Trapped in antlion holes: Through employing the sliding prey in the direction of antlion, ant is trapped in the antlion's pit. Alternatively, the movement of ant is surrounded by the location of antlion that could is modeled through altering the range of ant arbitrary walk in the direction of antlion location as in Eqs (4) and (5):

$$c^t{}_i = c^t + Antlion_j \quad (4)$$

$$d^t{}_i = d^t + Antlion_j \quad (5)$$

Here c^t is minimal of entire variables at generation t, d^t is maximal of entire variables at generation t, $c^t{}_i$ is minimal of entire variables for ant i, $d^t{}_i$ is maximal of entire variables for ant i, $Antlion_j$ determines the location of antlion j at generation t.

v) Arbitrary movements of ants: The arbitrary movements depends on Eq (6):

$$X(t) = [0, cumsum(2r(t1) - 1), cumsum(2r(t2) - 1), \ldots\ldots cumsum(2r(T) - 1)] \quad (6)$$

Here $cumsum$ evaluates cumulative sum, T is maximal amount of generations, t is the a single step for arbitrary movement, $r(t)$ is a stochastic function given as:

$$r(t) = \begin{cases} 1 \; if \; rand > 0.5 \\ 0 \; if \; rand \leq 0.5 \end{cases} \quad (7)$$

Here $rand$ is an arbitrary number obtained with uniform distribution over [0, 1]. To place arbitrary walks within the exploration domain, these are normalized with the help of Eq (8):

$$X_i^t = \frac{(X_i^t - a_i) \times (d_i - c_i^t)}{(d_i^t - c_i)} \times c_i \quad (8)$$

Here $a_i$is the minimal arbitrary movement for variable i, $d_i$is the maximal arbitrary movement for variable i , $c_i{}^t$ is the minimal of variable i at generation t, $d_i{}^t$ is the maximal of variable i at generation t.

vi) Elitism: To preserve the finest outcomes through the generations, elitism need to be employed. In this paper, the arbitrary movement of an ant is directed using chosen antlion and using elite antlion, and henceforth, the relocation of a specified ant follows the average of both arbitrary movements, as shown in Eq (9):

$$Ant_i = \frac{R^t{}_A + R^t{}_E}{2} \quad (9)$$

Here $R^t{}_A$ is the arbitrary movement round the antlion picked using the roulette wheel, $R^t{}_E$ is the arbitrary movement round the elite antlion.

4. Proposed feature selection approach using ALO algorithm

In this section, a semantic based text categorization approach is introduced that predominantly focus on the semantic senses of terms in the text document apart from the term frequency weight measure. The block diagram for the given approach is given in Fig 1. The sematic weights for the terms in the document represents the syntactic and semantic indices for the words reluctantly. Along with the sematic weight representation, the proposed approach also addressed the problem of 'curse of dimensionality' due to which the learning technique in text categorization has become a challenging task. Though there are numerous dimensionality reduction techniques by means of statistical and machine learning algorithms, in this paper, an intelligent nature inspired optimization algorithm is employed for feature selection. Similar to the traditional text categorization approaches, the proposed approach is also divided into three different phases. Each phase in the proposed approach has its own importance as the process of text categorization moves on. They are:

- Pre-processing phase
- Dimensionality reduction Phase
- Classification Phase

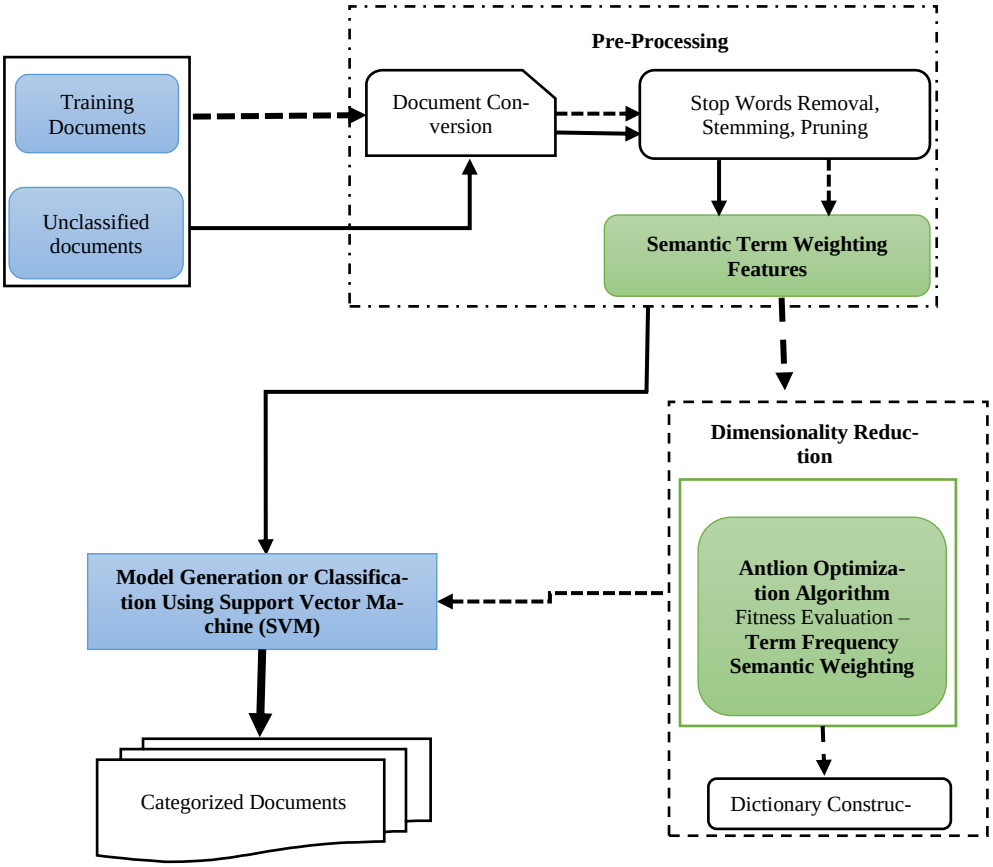


Fig. 1: Block Diagram of the Suggested Approach.

Pre-processing phase: Pre-processing eliminates the unwanted information from the text document. The most significant step in this phase is the representation of terms/words in text documents with its semantic weights.

a) Removal of stopping words: Words like conjunctions and pronouns which are not associated to notion of text known as stop-words. This procedure comprises of eliminating some common terms like 'a', 'an', 'the', etc., that comes frequently in entire textual documents. It is significant to eliminate these higher frequency terms since these might mis-categorize the documents. In this paper, stopping terms are eliminated in compliance with the prevailing stopping term list with 571 words.
b) Stemming: This procedure leaves the root forms of term. Thus, terms sharing similar root that appears to be diverse term because of affixes can be discovered.
c) Term Weighting: Once the terms are modified to its roots words, presentation form of document that is the expression have to be specified. This approach is known as weighting term. The usual form of term weighting is TF-IDF that depends on the notion that significance of any term to a document is based on its frequency along with the amount of exceptionality in the corpus. Thus, in this paper, a semantic term weighting structure is presented where weight of every term based on its semantic resemblance to the class.

The semantics of every class is defined using the senses of a set of terms present in the class label that are also interpreted by WordNet. The sense of each word is determined using general ambiguous framework based on optimization principal given in [5] known as Word Sense Disambiguation (WSD).

For any target word w with the collection of senses $S^w = \{s_1{}^w, \ldots\ldots, s_{ns(w)}{}^w\}$, the utmost probable sense of w present in the context $CW = \{w_1, \ldots..w_n\}$ is defined using:

$$Sense(w) = arg \max_{1 \le i \le ns(w_j)} \sum_{w_j \in CW} \max_{1 \le k \le ns(w_j)} sim(s_i{}^w, s_k{}^w) \quad (10)$$

Here $ns(.)$ denotes the number of senses for a word, and $sim(.,.)$ denotes the similarity measure of two senses. The hypothesis of employing maximization is that because the terms are appropriate to class, it is sensible to presume that sense of every term is the one with utmost identical to the semantics of class. Thus, maximization criterion is merely appropriate to top K terms, not to the collection of words in corpus. The Similarity measure employed is the Lin's similarity measure [6] depends on its theoretical basis and its higher performance:

$$sim^{Lin}(s_1, s_2) = \frac{2\log(p(LCA(s_1,s_2)))}{\log(p(s_1))+\log(p(s_2))} \quad (11)$$

Here $LCA(s_1, s_2)$ denotes to the lower mutual ancestor of senses s_1 and s_2 in the hierarchy of senses. $sim^{Lin}(s_1, s_2)$ ranges amongst 0 and 1. For any word w with a group of senses $S^w = \{s_1{}^w, \ldots\ldots, s_{ns(w)}{}^w\}$, it is given as:

$$S_L(i) = \sum_{w_j \in W_L{}^c} \max_{1 \le k \le ns(w_j)} sim(s_i{}^w, s_k{}^w) \quad (12)$$

And

$$S_N(i) = \sum_{w_j \in {W_N}^c} \max_{1 \le k \le ns(w_j)} sim(s_i^w, s_k^w) \quad (13)$$

Where ${W_L}^c$ is the collection of words present in the label C of each category and ${W_N}^c$ is the collection of words appearing in the WordNet interpretation of every term in ${W_L}^c$. Then the sense of the word w is defined using:

$$Sense(w) = arg \max_{1 \le i \le ns(w_j)} (T(S_L(i), 0, -\infty) + S_N(i)) \quad (14)$$

Where

$$T(\chi, \delta, \gamma) = \begin{cases} \chi & if\ \chi > \delta \\ \gamma & if\ \gamma \le \delta \end{cases}$$

d) Pruning of the words: This approach fundamentally filters lesser repeated features in a document set. The term vector is very higher magnitude and sparse. Further, it is observed that numerous elements in vector is ''0''. Thus, the pruning of terms occurs lesser than two times in the documents. This procedure minimizes the term vector dimension.

Dimensionality Reduction Phase: This is the utmost significant phase in text categorization which influences the accuracy of documents through text categorization. Though the unwanted and additional information are pruned using pre-processing phase, there still remain a high dimensional feature with noisy and irrelevant information. These inappropriate and recurrent features frequently reduce the efficiency of categorizers in speed and accuracy. In this paper, antlion optimization algorithm is employed to perform feature selection as to minimize the relevant data loss and increase the minimization of inappropriate features. The fitness evaluation employed in this algorithm is the Term Frequency Semantic Weighting approach.

Antlion Optimization Algorithm for Feature Selection: The working of this algorithm is given in terms of ant and antlion working operations. The operation in this algorithm involves the following steps.

1) Arbitrarily initialize the population of ant location Ant and the population of Antlions position $Antlions$ with the m features present in n documents in a Matrix of dimension$m \times n$.
2) Estimate the fitness of entire ants and antlions using fitness function as given below:

$$TFSW(w,d) = count(w,d)(s(w,C) + \theta \propto count(w,d)(\frac{s(w,C)}{\theta} + 1) \quad (15)$$

Here, $TFSW(w,d)$ represents the Term Frequency Semantic Weighting scheme (TFSW) for a word w in document d pertaining to category C, θ serves to be smoothing factor for words

3) Find the Fittest Antlion from the evaluated fitness values as Elite.
4) Initialize t=0 and while $(t \le T)$
a) For every Ant_i i.e. document
i) Choose an Antlion employing Roulette wheel selection operation
ii) Slide the ant i.e. features in each document toward the antlion i.e. the selected feature in the Antlion using equation (2).
iii) Generate an arbitrary movement for Ant_i and normalize it using equations (4), (5) for modeling trap, equation (6) for arbitrary movement and equation (8) for walk normalization.
b) End
c) Re-evaluate the fitness of complete ants i.e. documents.
d) Replace the antlion (selected feature) with its corresponding ant (best obtained feature in document) if the ant becomes fitter using the equation (1).
e) Update the Elite if an antlion is fitter compared to the present elite
f) $t = t + 1$
5) End while.

Classification Phase: The SVM is employed for classifying the reduced text file document comprising of merely the relevant and significant features for classification. SVMs attain extensive enhancements over the presently finest performing approach and works strongly over a diverse learning tasks. SVMs are a usually appropriate device for machine learning. Presume that the training instances x_i, and target values$y_i = \{-1,1\}$ are given. SVM explores for a separating hyper plane that divides positive and negative instances from one another using maximum margin, alternatively, the distance between the decision surface and the nearest instance example is maximum.

SVM depends on Structural Risk Minimization framework [9] obtained from computational learning model. The notion of structural threat reduction is to discover a hypothesis h for that could assure the lower most true error. The true error of h is the possibility that h would make an error on a hidden and arbitrarily chosen test instances. A higher bound is employed to link true error of a hypothesis h with error of h on training set and complexity of H, the hypothetical domain comprising h [9].

5. Experimental results and its analysis

The experimental results for suggested Antlion Optimization based Text Categorization is carried out using two different data samples.

i) Reuters-21578: This data set4 comprises of 21578 articles obtained from Reuter's newswire and was downloaded from the web site, https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection. After eliminating documents deprived of labels or deprived of body, the training data samples comprises of 7775 text files, and the test data samples comprises of 3019 text files.

ii) 20 Newsgroup Letters: This dataset was obtained from http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html. In the collection, completely, 20,000 news articles and 20 classes are present. The 20 classes are specified as the bi-level hierarchical structure; the initial level has four classes and subsequent level has 20 classes in turn every class in the initial level has five classes in the subsequent level. In this paper, four categories are chosen. They are: computer, record, natural science, and social science. In every class, the training set comprises of 2,800 news articles in the temporal order, 700 news articles per a category, where the test set contains 1,200 news articles, 300 news articles per a category.

5.1. Performance metrics

In most text classification, the performance of feature selection techniques is particularly important. Several norms like precision and recall are often employed to estimate the efficiency of feature selection approach.

- Precision is described as the ratio of accurate cases to entire predicted cases. Assume that TP_i represents the number of test documents correctly categorized under i^{th} category C_i and FP_i denotes the number of test documents erroneously categorized C_i; then classification precision can be formulated as

$$Precision = \frac{TP_i}{TP_i + FP_i} \quad (16)$$

- Recall is described as fraction of the correct cases to the total cases. Assume that TP represents the number of test documents correctly classified under i^{th} category C_i, and FP_i is the number of test documents falsely categorized under other categories; these possibilities might be evaluated in terms of the contingency table for C_i; then classification recall can be formulated as

$$Recall = \frac{TP_i}{TP_i + FN_i} \quad (17)$$

5.2. Experimental results analysis

To analyses the performance of feature selection based text categorization, the proposed Antlion Optimization Algorithm based Feature Selection (AO-FS) is compared with the existing feature selection text categorization approaches such as Table Based Matching using Genetic Algorithm (TBM-GA), Semantic Weighting based Genetic Algorithm (SW-GA) and Sematic Weighting based Particle Swarm Optimization (SW-PSO). Figures from fig 2 to fig. 2 to fig. 6 illustrates the performance of the suggested AO-FS methodology against the TBM-GA, SW-GA and SW-PSO for the ten most recurrent categories in contradiction to the classification accuracy (precision & recall).

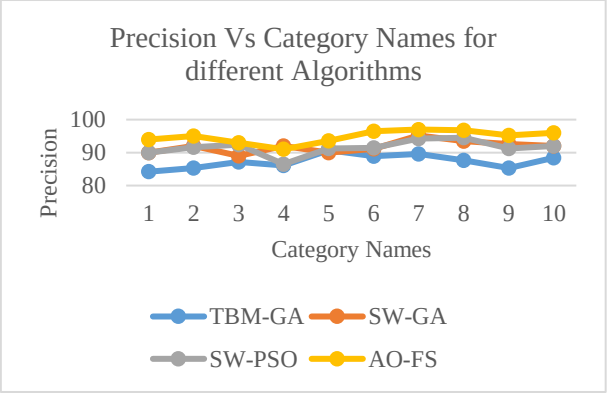


Fig. 2: Precision vs. Category Names for Different Algorithms.

Fig. 2 represents the precision value of TBM-GS, SW-GA, SW-PSO and proposed AO-FS with dissimilar no. of categories. Fig. 3 represents the recall value of TBM-GS, SW-GA, SW-PSO and proposed AO-FS with dissimilar no. of categories. Fig. 4 represents the precision value of TBM-GS, SW-GA, SW-PSO and proposed AO-FS with dissimilar no. of features. Fig. 5 represents the recall value of TBM-GS, SW-GA, SW-PSO and proposed AO-FS with dissimilar no. of features. Fig. 6 represents the average fitness value obtained by the algorithms such as GA, PSO and Antlion Optimization Algorithm.

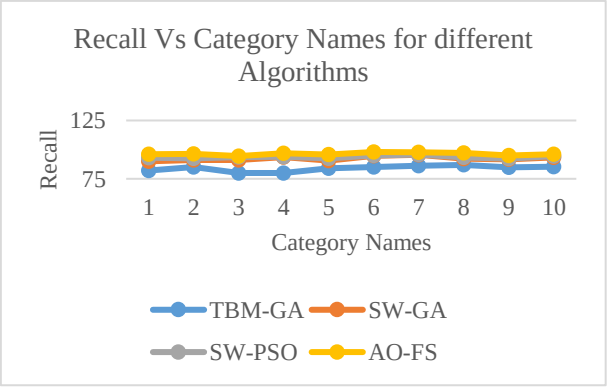


Fig. 3: Recall vs. Category Names for Different Algorithms.

From the experimental result in Fig. 2, it can be seen that the Antlion Optimization Algorithm based Text Categorization has the highest precision value for different categories and its maximum value is close to 96%. However, the precision to 84%. From the experimental results in Fig. 3, it is witnessed that AO-FS is significantly better compared to the other three prevailing approaches pertaining to the recall value. The maximum recall value of AO-FS is 98%. Nevertheless, the recall of TBM-GA is comparatively minimum and its minimum value is 76%. From the experimental results in Fig. 2 and Fig. 3, it can be flexibly inferred that AO-FS Approach attained better performance with reduced feature set compared to other two algorithms, particularly in recall.

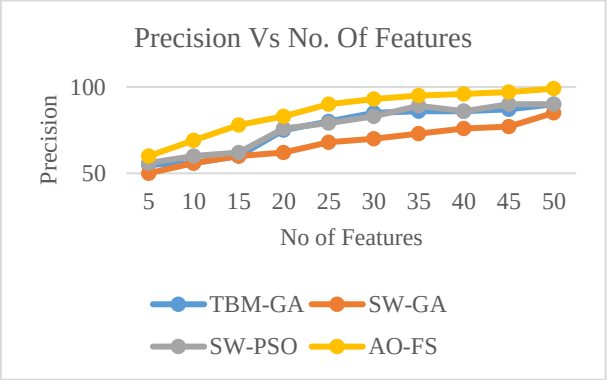


Fig. 4: Precision vs. No. of Features for Different Algorithms.

When a gradual increase is in the number of features, the precision and recall of the three feature selection algorithms are gradually increased. As observed in Fig. 4 and Fig. 5, the overall performance of AO-FS is significantly superior to TBM-GA, SW-GA and SW-PSO. It is worth considering that the suggested approach has the least number of support vectors compared with other feature selection approaches. From the experimental outcomes in Fig, 6, the average fitness of AO-FS is the highest in most cases and its maximum value is close to 0.81. It could be seen from the experimental outcomes that CGFSO learning process effectively and efficiently reduces the complexity of the system in the feature selection stage.

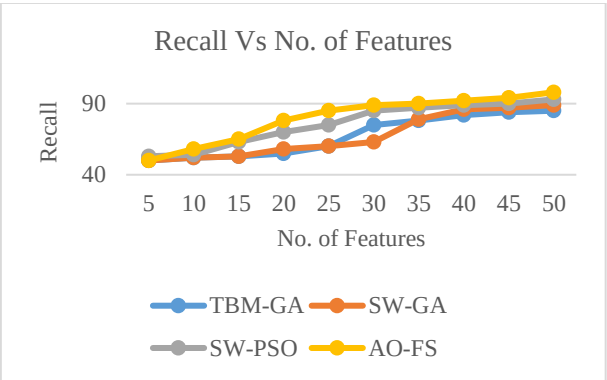


Fig. 5: Precision vs. No. of Features for Different Algorithms.

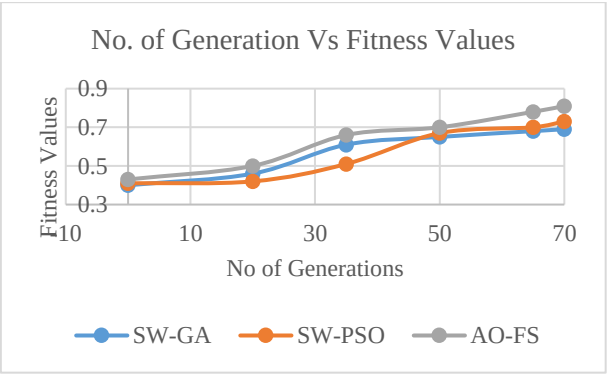


Fig. 6: Fitness Values vs. No of Generation for Different Algorithms.

6. Conclusions

Pertaining to the Big Data era, there is an extensive rise in textual information, text categorization has become a way to process and organize the text data. Text categorization denotes to the approach of classifying a text into its category or categories amongst the predefined ones. Its preliminary task is to predefine a list of categories and allocate sample texts each of them. So as to achieve the goal of reduced dimension, in this paper, Antlion Optimization Approach based Feature Selection approach is suggested of accurate and efficient text categorization. Antlion Optimization Algorithm is a meta-heuristicbio-inspired that mimics the nature of ant and antlion and its interaction with each other. The experimental results also illustrated that the proposed Antlion Optimization Algorithm based feature selec-

tion has exhibited better performance compared to the existing meta-heuristics approaches such as genetic approach and particle swarm optimization. The precision and recall is also maximum when compared to the existing approaches.

References

[1] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Application of ant colony optimization for feature selection in text categorization," in Proceedings of the IEEE Congress on Evolutionary Computation (CEC '08), pp. 2867–2873, IEEE Press,HongKong, June 2008.https://doi.org/10.1109/CEC.2008.4631182.

[2] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", pp839-849, Soft Computing, Vol 19, No 4, 2015.https://doi.org/10.1007/s00500-014-1411-9.

[3] Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. 14th Int. Conf. on Machine Learning ICML-97. (1997) 412–420.

[4] Caropreso, M., Matwin, S., Sebastiani, F.: A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: Text Databases and Document Management: Theory and Practice. Idea Group Publishing, Hershey, PA (2001) 78–102

[5] Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proc. 22nd Int. ACM SIGIR Conf. on R. & D. in Information Retrieval. (1999) 42–49.https://doi.org/10.1145/312624.312647.

[6] Mladeni´c, D.: Feature subset selection using in text learning. In: 10th European Conference on Machine Learning. (1998) 95–100https://doi.org/10.1007/BFb0026677.

[7] Z. Zhen, X. Zeng, H. Wang, and L. Han, "A global evaluation criterion for feature selection in text categorization using Kullback-Leibler divergence," in Proceedings of the International Conference of Soft Computing and Pattern Recognition (SoCPaR '11), pp. 440–445, October 2011.https://doi.org/10.1109/SoCPaR.2011.6089284.

[8] Ramakrishna Murty, M., Murthy, J., Prasad Reddy, P., Satapathy, S.: A survey of cross domain text categorization techniques. In: Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, IEEE (2012) 499–504https://doi.org/10.1109/RAIT.2012.6194629.

[9] Nguyen, C.T.: Bridging semantic gaps in information retrieval: Context-based approaches. ACM VLDB 10 (2010).

[10] Rafi, M., Hassan, S., Shaikh, M.S.: Content-based text categorization using wikitology. CoRR abs/1208.3623 (2012)

[11] Forman, G.: An experimental study of feature selection metrics for text categorization. Journal of Machine Learning Research 3 (2003) 1289–1305.

[12] K. Waqas, R. Baig, and S. Ali, "Feature subset selection using multi-objective genetic algorithms," in Proceedings of the 13th IEEE International Multitopic Conference (INMIC '09), pp. 1–6, December 2009.https://doi.org/10.1109/INMIC.2009.5383159.

[13] A. AlSukker, R. N. Khushaba, and A. Al-Ani, "Enhancing the diversity of genetic algorithm for improved feature selection," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '10), pp. 1325–1331, October 2010.https://doi.org/10.1109/ICSMC.2010.5642445.

[14] M. Mahrooghy, H. Y. Nicolas, G. A. Valentine, A. James, and Y. Shantia, "On the use of the genetic algorithmfilter-based feature selection technique for satellite precipitation estimation," IEEE Geoscience and Remote Sensing Letters, vol. 9, no. 5, pp. 963–967, 2012.https://doi.org/10.1109/LGRS.2012.2187513.

[15] Hao Chen, Wen Jiang, Canbing Li, and Rui Li, "A Heuristic Feature Selection Approach for Text Categorization by Using Chaos Optimization and Genetic Algorithm", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2013, Article ID 524017, 6 pages.https://doi.org/10.1155/2013/524017.