ECO INSPIRED BEES: A NOVEL FEATURE SELECTION MECHANISM FOR SENTIMENT ANALYSIS

R. Sharmila¹, M. Sivajothi^{2 §}

^{1,2}Manonmaniam Sundaranar University

Tirunelveli, INDIA

¹Department of MCA

Er. Perumal Manimekalai College of Engineering

Hosur, INDIA

²Sri Parasakthi Collere fog Women

Courtallam, INDIA

Abstract: Sentiment analysis also called as opinion mining is defined as the programmed learning of public's opinion, evaluations, and sentiment on different events, objects as well as their respective properties. Recently, it created high impact in academia and industry due to number of growing research problems and various applications. Selecting the features play a vital role in supreme data mining tasks since it assists in the reduction of data dimensionality by leaving the non-relevant features behind. Present work applies the ecological concept of habitat, ecological relationship and ecological succession for building an Ecology-Inspired Optimization algorithm named as Ecology Inspired Artificial Bee Colony algorithm (EI-ABC). Proposed method uses various population of candidate solution which cooperates and co-evolves with one another, as per the given meta-heuristic algorithm. Sentiments are classified by using two classifiers viz., K Nearest Neighbor (KNN) and Classification and Regression Tree (CART). Result suggests that the EI-ABC algorithm proves to be interesting alternative for numerical optimization.

AMS Subject Classification: 62M10, 05C05, 92B20

Key Words: opinion mining, sentiment analysis, feature selection, Ecology-Inspired Optimization (ECO), Artificial Bee Colony (ABC), K-Nearest Neighbor (KNN), Classification and Regression Tree (CART)

Received: February 10, 2017

Revised: April 4, 2017

Published: May 6, 2017

© 2017 Academic Publications, Ltd.

url: www.acadpubl.eu

[§]Correspondence author

1. Introduction

1.1. Opinion mining

Opinion mining refers to the process of automated knowledge extraction from the public opinion on a specific topic or issue. Human thought as well as user's opinion has high potential towards discovery of knowledge and decision support. The objective of opinion mining is to enable computer to identify and express emotion. A perception, thought, or attitude on an emotion instead of reasons is called sentiments. Hence opinion mining is also called as sentiment analysis. Most of us expect help from others for decision making because it is believed that best decisions are only based on others opinion. Surveys, blogs and review sites are utilized for collecting customer's opinion on various products or service to obtain the knowledge on the company reputation of the market [1].

Opinion mining is a Natural Language Processing (NLP) and Information Extraction (IE) task that has aim for obtaining emotions of the writer expressed as positive or negative comment through the analysis of numerous documents. It merges the technique of computational linguistic and Information Retrieval (IR). The major aim of Sentiment analysis the classification of documents and find the polarity. Polarity can be expressed as positive, negative or neutral. Sentiment analysis can be done in three levels [2]: Document level: Classifies the entire document as positive, negative or neutral. Sentence level: Classifies the complete sentence as positive, negative or neutral. Aspect & Feature level: Classifies sentence/document as per the aspect of it as positive, negative or neutral.

Machine learning [3] algorithm enhances the performance with experience by default. Prediction is the expected main result of the enhanced performance. Any algorithm is proved to be learnt if it enhances its capability in predicting key element of any task while providing accurate data. Machine learning algorithm is featured by the language that represents knowledge. Researches reveal that single learning approaches are not superior and normally various learning algorithm produces the similar result. But the data's nature that features the task as learned has more impact on the success of learning algorithm. Learning becomes failure in machine learning algorithm if data do not present statistical regularity. Even though new data can be built from old one for exhibiting statistical regularity and hence learning could be facilitated, but it seems to be a difficult task that a complete automatic system is complex.

Feature refers to an individual quantifiable property of the process that is being considered. Using a collection of feature, machine learning algorithms

will be able to classify. Various techniques are designed for addressing the challenge to reduce non-relevant and redundant variables that acts as trouble for completing the task. Feature selection, otherwise called as variable elimination assists to understand the data, reduce computation need, and reduces the impact of curse of dimensionality as well as improves the performance of the predictor. Feature selection's objective focuses on selecting subset of variable from the given input that will effectively explain the input data during reducing effect from noises or non-relevant variable but yet provides optimal prediction result [4].

1.2. Feature Selection

Three categories of feature selection methods are i.e. filter, wrapper and embedded. In filter categories, a collection of features are chosen according to a specific mathematical equation and might be applied with a classifier. In contra to this, the feature selected in the techniques of wrapper and embedded, they are attached with specific classifiers. Except that it is rigid in terms of classifiers, wrapper and embedded usually needs huge resource allocation and requires more time for execution [5].

Document Frequency (DF), Chi-Squared (CHI) and Information Gain (IG) are few instances of feature selection methods that falls under the filter technique. DF calculates the frequency of a specific term in all methods. Feature that has most and least frequency are eliminated. CHI estimates the degree of a specific feature that is not related to that specific class. Feature is arranged as per its relevancy with a specific class. IG estimates the relevancy of a feature as per the probability of the existence of a word in the specific class.

Four types of feature categories are used for sentiment analysis. Syntactic feature use words/ Part of Speech (POS) tags, N-gram, phrase pattern or punctuations, one among all. The cited author observed that phrases pattern like "n+a_j" (noun succeeded by positive adjectives) classically represents positive sentiments orientation, whereas "n+d_j" (noun succeeded by negative adjectives) mostly exhibits negative sentiments. Semantics focus on the relationship among signifier like word, phrase, sign and symbol. Linguistic semantic helps understanding human expressions via languages. Score-based technique is normally used to conjunct with semantic features. This technique normally classifies message sentiments according to the title summation of included positive and negative semantics feature [6].

Link based sample is categorised with the relation and link available between them. Link-based feature uses links/citations analysis for determining

sentiment for Web artifact and document. Stylistic feature that is used by artist in the trial to send messages to us. Very little use of stylistic feature like word-length distribution, vocabulary richness measure, character- and word-level lexical feature, and special- character frequency are found. Stylistic feature uncovers latent pattern that improves the performance of sentiment classification.

Feature selection comprises four phases' viz., subsets generation, subsets evaluation, stopping criteria and results validation. Subsets generation means search process for evaluating candidate feature subset according to a particular strategy of searching. Candidate subset is assessed and compared with previous best known particular evaluating criterion. The best and novel subset substitutes the preceding subset of same quality. This subsets generation and assessment step is performed repeatedly till a terminating condition is met. The best subset selected is later tested by previous knowledge or various validating procedures by using manipulated and/or realistic data set [7].

1.3. Motivation and Justification

In recent day's social media is the optimal tool to understand the people's opinions, advices, comments, complements and their perceptions towards any product policies and services. Automated detecting of sentiments from given opinion is becoming significant from application's perception. Such opinion mining or sentiment analysis was done using various machine learning algorithms. Motivated with the demanding trend for proper classification of sentiments, and the performance of swarm intelligence stimulated algorithms in classification, it is proposed to develop an enhanced ABC named as Ecology Inspired Artificial Bee Colony algorithm (EI-ABC), combined with KNN and CART classifier.

1.4. Outline of the Proposed Work

This work proposes EI-ABC algorithm for the feature selection for sentiment classification, while KNN and CART are used as classifiers. The features are selected given sentiment/opinion from the given data set by using ABC with KNN, ABC with CART and EI-ABC with KNN and EI-ABC with CART. The outline of the proposed work is shown in Fig.1

Figure 1: Outline of the proposed work

1.5. Organization of the Paper

The remainder of the work is structured as: in Section 2, related works in literature are discussed. In Section 3, the materials and methods used in the presented work is explained. Section 4, discusses about the results, concluded in Section 5.

2. Related Works

Winkler et al., [8] proposed a collection modeling method for sentiment analysis with machine learning algorithm that depends on the word analysis found into the sentence and the establishment of huge set of heterogeneous model, i.e., binary and multi-class classification model that is estimated by several distinct machine learning techniques; this model represents the relationship among the existence of given words (or group of words) and opinions/sentiments. The result attained using a German corpus of Amazon recession and a group of machine learning techniques (decision trees and adaptive boosting, Gaussian processes, random forests, KNN classification, Support vector Machines (SVM) and Artificial Neural Networks (ANN) with evolutionary features and parameters optimization, and genetic programming). Duve & Seth [9] presented and enhanced the approaches to classify the reviews and also for detecting the review polarity with machine learning technique. ABC algorithm is applied to classify text in three categories: negative, positive and neutral. In the results, nature inspired ABC classifier BON proved to provide optimal results compared to BOW, SVM with both BON and BOW.

Tripathi & Nangana [10] demonstrated a paradigm for analysing movie review sentiments using the merged technique for NLP and machine learning. The behaviours of two classifiers, Naive Bayes and SVM, are studied by combining with various feature selection approaches for obtaining the result of sentiment analysis. Finally, the intended model for analysing sentiment was applied for obtaining the result for high order n-grams. The sentiment analysis

is used in health care systems. Gopalakrishnan et al., [11] presented an opinion mining dimensionality reducing method for mining users developed health review sets. The novel techniques classify patients review from online forum as positive/negative by default. Result shows the novel dimensionality reducing techniques are efficient in classifying. Manek et al., [12] presented a Gini index oriented feature selection technique with SVM classifiers to classify the sentiments of large movie review data set. The result shows that the Gini index technique has optimum classification performance as reduced rate of error and precision.

Sharma & Dey [13] provided a hybrid sentiment classification paradigm depending on boosted SVM. The intended paradigm model exploits the performance of classification of two methods (Boosting and SVM) used for sentiment based online review classification. Specifically, the results show that SVM ensemble with bagging or boosting significantly outperforms a single SVM in terms of accuracy of sentiment based classification. Agarwal & Mittal [14] proposed the unigram and bi-gram that are extorted from texts, and composite feature is generated using them. IG and Minimum Redundancy Maximum Relevancy (mRMR) feature selection techniques are applied for extracting prominent feature. Moreover, impact of several feature set for sentiment classifications were studied by the use of machine learning techniques. Effect of various category of features are studied on four standard datasets i.e. Movie reviews, product (book, DVD and electronics) reviews dataset. Empirical result shows that composite feature generated from important feature of unigrams and bi-grams outperform sentiment classification when compared to other features.

Agarwal et al., [15] presented the bi-tagged phrases are used as features in combination with unigram features for sentiment classification. Therefore, a feature selection method was used to select only relevant features from the feature vector. Experimental results show that the combination of prominent unigrams and bi-tagged phrases outperforms other features for sentiment classification in a movie review dataset. Agarwal & Mittal [16] proposed the BoW techniques that assume the independence of word and ignore the significance of semantic and subjective information present in the text. Machine learning algorithm reduces this high-dimensional features space by using feature selection technique that chooses only significant features by removing the noisy and non-relevant feature. In recent trend, machine learning oriented sentiment analyzing models are receiving prominent interest in the field. Ahmad et al., [17] proposed the feature selection methods in sentiment analysis on the basis of NLP and latest techniques like Genetic Algorithm (GA) and Rough Set Theory (RST). This research compares feature selections in text classification oriented

on conventional and sentiment analysis techniques. Finally it is concluded that meta heuristic oriented algorithm has the potentiality that if applied in sentiment analyzing researches, will be able to produce a best subset of features by removing the redundant and non-relevant features.

3. Methodology

Original subset of features is chosen by feature selection technique, and the optimality of feature subset is assessed by evaluating criteria. If domain dimensionality extends, N feature count improves. This is inflexible to find optimal features subsets. Feature selection associated problem is NP-hard. Machine learning feature selection means a global optimization issue that minimizes feature quantity, removes non-related, noisy and redundant data that results in recognizing accurateness. This is a significant phase that affects the performance of pattern recognition. Normally, feature selection challenges are solved by using a single objective optimization technique. This work uses, IMDB dataset, ABC and EI-ABC based feature selection algorithm, KNN and CART classifier for feature selection and the same are described.

3.1. Feature Selection Algorithm

3.1.1. Artificial Bee Colony (ABC) Algorithm

Feature selection minimizes the feature space dimensionality for restricting the storage requirement and increases algorithm's speed; it eliminates the data that is more redundant, non-relevant or noisy. The immediate effect for data analysis task is increasing the speed of executing time of learning algorithm and improves the quality of the data, increases the precision of the final paradigm. Feature set reducing saves resource in the succeeding round of data gathering or while utilizing; Performance enhancement, for gaining predictive accurateness; Data understanding for gaining knowledge regarding the process that developed the data or just for visualizing the data [19].

The ABC algorithm is said to be a swarm based, meta-heuristic algorithm that is motivated by the foraging behaviour of honey bees with their colonies. The paradigm comprises three vital components: employed forager, unemployed forager, and food source that is near the hive. Two different leading behaviour modes are explained with this model. They are essential for self-organizing and collective intelligence: recruiting forager bee to a rich food source, results into

positive feedback whereas, abandoning poor source by a forager causes negative feedback [20].

The ABC includes three categories of artificial bee: employed forager, onlooker and scout. The employed bee comprises the primary part of the colony and onlookers will be the secondary part. The employed bee is associated to the specific food source. Or otherwise, the quantity of employed bee is same as the quantity of food source of the hive. The onlooker observes the employed bee's dance in the dancing area for selecting the food source, and the scout searches new food source in random. Similar to the optimization framework, the quantity of food source (means the employed bee or onlooker bee) in ABC algorithm is equal to the quantity of solutions available in the population. And more over, the food source position signifies the promising solution's position towards the optimization issue, where the nectar's quality represents the solution's fitness cost (quality).

In ABC algorithm imitation with the bee behaviour depends on a simple model comprising three necessary components: food source, worker bee (employee), and non-worker bee (onlooker and scout). The food source represents the solution of an issue, and bee is the solutions searching agent. The input parameters are: Colony Size (CS), Number of Design Variable (D), the Maximum Cycle Number (MCN), and the limit for withdrawal from a source (Limit). In this algorithm, the complete population of the colony of bees (CS) is sub-divided equally among the employee and onlooker. The first source distribution is generated in random in the form (1) [21]:

$$x_{ij} = x_{min,j} + rand(0,1)(x_{max,j} - x_{min,j}); \quad i = 1, ..., CS/2; \quad j = 1, ..., D \quad (1)$$

Where, $rand(0,1)$ represents a random number among 0 and 1.

Every employee is given to a source (x_i) and it develops a candidate solution (v_i) commencing the process of exploitation (local search), in the form (2):

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}); \quad j \in 1, ..., D; \quad k \in 1, ..., CS/2; \quad k \neq i \quad (2)$$

Where, represents random number between -1 and 1.

Next, it chooses the best among x_i and v_i . The assessment of the source depends on its fitness value (fit_i) related to the objective function (f_i) in the form (3):

$$fit_i = \begin{cases} 1/(1 + f_i) & \text{if } f_i \geq 0 \\ 1/abs(f_i) & \text{if } f_i \leq 0 \end{cases} \quad (3)$$

If the solution does not show any improvement in a number of trials according with the parameter limit the employee bee becomes a scout bee which randomly replaces the source, according to equation (1). For each current source (x_i), it is calculated the recruitment probability value (p_i) using equation (4). Thus the onlooker bees can be probabilistically recruited for the most promising sources. In the natural behavior of bees this step is known as the "waggle dance".

$$p_i = \frac{fit_i}{\sum_n^{SN} fit_n} \quad (4)$$

The onlooker bee recruited for a given source becomes an employee bee and the iterative cycle of the algorithm starts again [22]. The pseudo-code of the basic algorithm is shown as follows.

DataInputCS, Limit_{max}, MCN, D, [x_{j,min}; x_{j,max}] $\forall j = 1, \dots, D$.

Initialization Cycle $g = 1$;

Initialsolutionsdistribution $x_i^0 \quad \forall i = 1, \dots, CS/2$.

While $g \leq MCN$

Generatesmodifications v_i^g ;

If improved then $x_i^g = v_i^g$ *and*

Limit = 0; *else*, *Limit* = *Limit* + 1

Calculateprobabilities p_i^g

Probabilisticselectionofthemost promising x_i^g *solutions*;

If $p_i^g \geq \text{rand}(0, 1)$ *recruitsonlooker*

ReplacessolutionsinwhichLimitreached

thecountandmakesLimit = 0

Memorizesbestsolutionfoundsofar

$g = g + 1$

Repeat

End

3.1.2. Proposed EI-ABC Algorithm for Feature Selection

The EI-ABC algorithm presents a novel view for developing cooperative searching algorithm. The EI-ABC comprises population of individuals (candidate solution towards a problem to be addressed) and every population arises as per a searching strategy. Similarly, the individual of every population is updated based on the mechanism of amplification and diversification, and initial parameter, particular to the searching strategy that is used. The eco-inspired

model behaves in two methods: homogeneous or heterogeneous. A homogeneous behaviour states that all population evolves with regard to the similar optimization technique, configured with similar parameter. A modification in the strategy or parameter in minimum one population features heterogeneous behaviour.

The ecological inspiration spreads by the application of certain ecological concept like: habitat, ecological relationship and ecological succession. If distributed in the searching space, population of individuals developed in the similar area constitutes ecological habitats. Hence habitats are groups whose components are population that belongs to the similar area of the searching space. For instance, in a multi modal hyper-surface, every peak becomes a potential habitat for few populations. A hyper-surface possesses different habitat and naturally, the population moves around through complete environment. Anyhow, every population belongs only to single habitat in given time t . Hence, by definition, the intersection among the habitats at time t is the empty set. Two types of ecological relationship are defined with the habitat definition. Intra-habitat relationship occurs between population within every habitat and inter-habitat relationship occurs between habitats.

In EI-ABC, intra-habitat relationships refer to the mating of individuals. Population belonging to the single habitat develops a reproductive relation among their individuals, mixing the population and favours the co-generation of the included population. Population belonging to several habitats are known as separated reproductively. The inter-habitat relationships are the higher migration. Individuals of said habitat migrate to other habitat with the focus to identify potential region for surviving and mating [23].

Along with the mechanism of amplification and diversification in particular to every searching strategy, while taking into account the proposed algorithm's ecological context, the intra-habitat relationship is in charge to intensify the search, whereas the inter-habitat relationship is incharge to diversify the search. Within the ecological metaphors, the ecological succession represents the transformational procedure of the system where the populational group is formed (habitat); relationship of population is established along with stabilizing of the system by self-organizing its elements.

Pseudo code for EI-ABC is follows:

```

1 : Consider  $i = 1, \dots, NQ$ ,  $j = 1, \dots, NH$  and  $t = 0$ ;
2 : Initialization each population  $Q_i(t)$  with  $n_i$ 
   random candidate solutions;
3 : while stop criteria not satisfied do

```

Ecological succession cycles

- 4 : *Perform evolutive period for each population $Q_i(t)$;*
- 5 : *Apply metric C_i to identify the region of reference
for each population $Q_i(t)$;*
- 6 : *Using the C_i values, define the NH habitats;*
- 7 : *For each habitat $H_j(t)$ define the communication
topology $CT_j(t)$ between populations $Q_i^j(t)$;*
- 8 : *For each topology $CT_j(t)$, perform interactions
between populations $Q_i^j(t)$;*
- 9 : *Define communication topology $TH(t)$
between $H_j(t)$ habitants;*
- 10 : *For $TH(t)$ topology, perform interactions
between $H_j(t)$ habitants;*
- 11 : *Increase t*
- 12 : *end while*

The pseudo-code is explained as follows: initially at time $t = 0$, every individual of entire population Q_i , with $i = 1 \dots NQ$, are arbitrarily initialized (line 2). NQ is the total quantity of population. The major loop states the ecological succession (line 3 to 12). The primary step within the major loop refers to the evaluative duration for all populations $Q_i(t)$ (line 4). In this step every population searches the search area without obstacle based on the amplification and diversification criterion.

In the termination of evaluative duration of entire population it is essential to recognize the regions of reference for every population (line 5). The regions of reference are applied for defining the system's habitat. The measure selected for defining the regions of reference is the centroid C_i , with $i = 1 \dots NQ$. Centroid denotes a point in the space where longest concentration of individual of population i exists. If the centroids are found, then the distance between population is estimated and the habitat $H_j(t)$ is defined with a minimum distance threshold $\rho \in [0 \dots 1]$ (line 6). The populations distance is based on the definitions and representations of every population. Here Euclidean distance is used. At last entire population is linked with a habitat: $H_i^j(t)$, population i belong to the habitat j at time t .

By defining NH habitat, it is essential to finalize the intra-habitat communication topology $CT_j(t)$ for every habitat $H_j(t)$ with $j = 1 \dots NH$ (line 7). The topology is defined by an adjacency matrix among the population that belongs to every habitat. After defining the topology of intra-habitat, then the mating ecologic relationship among adjacent population occur (line 8). Here,

single individual of every population is selected by the tournament strategies and genetic exchanges are done among them for generating a fresh individual. That individual substitutes an individual chosen randomly from its population, except the optimal individual.

After establishing the interaction among the population of every habitat, the TH(t) interaction topology between habitats (line 9) is arbitrarily defined. This inter-habitat topologies TH(t) are taken to complete the huge migration ecologic relationships. Here, in these relationships, for every habitat an arbitrary population that belongs to that is selected. The optimal individual of these populations migrate to other habitats and at the destination habitat, it substitutes an individual that is chosen in random, except the optimal individual (line 10). The major loop goes on till the ecological successions cycle reaches the maximum predefined value.

3.2. Classifiers

3.2.1. k-Nearest Neighbours (KNN)

Classification [24] is the challenge to identify to which group of individuals (sub-populations) a novel observation belongs to based on training data set having observation information and whose categories are known. Most of the realistic problems could be paradigm as classification challenges like noting a given mail as "spam" or "non-spam" class, automatically assigning the categories (e.g., "Sports" and "Entertainment") of forthcoming news, and assignment of diagnosis for a given patient as explained by noticed characteristic of the patients (Gender, BP, Some symptoms, etc.).

Nearest neighbour techniques are taken as the most easiest but efficient class of existing classification algorithm. Their standard depends on the assumption that, for a given collection of instance stored in a training data set, the classes of a novel yet not observed occurrences is mostly to be the majority of its nearest "neighbour" instance available in the training data set. Hence the KNN algorithm executes by verifying the k closest instance available in the data set to novel occurrences that need to be categorized, and predicts depending on which class the most of the k neighbour belongs to. The concept of nearness is usually given with the distance function among two points in the space attribute, specified priori as parameters to the algorithm. Instance of distance function normally typically applied is the standard Euclidean distance between two points in an n-dimensional space, where n is the number of attributes in the data set [25].

For classifying a review, the KNN classifier approximately ranks the review from the training review set, prior to the classification based on the k most identical neighbours. While presenting with test review d , the classifier locates the KNN from training review. The measure of all nearest neighbour reviews that is mostly identical to the test reviews are used as the weight of the neighbour reviews class [26].

The weighted sum in KNN classification can be represented as follows equation (5):

$$score(d, t_i) = \sum_{d_i = KNN(d)} sim(d, d_j) \delta(d_j, c_i) \quad (5)$$

Where, $KNN(d)$ represents the collection of K-Nearest Neighbour for the reviews (d). If d_j is belonging to c_i , then $\delta(d_j, c_i)$ can be either equal to 1 or 0. In case of test review, d shall be fitted into the class that has highest resulted weighted sum.

3.2.2. Classification and Regression Tree (CART)

CART is an iterative partitioning technique applied for predicting continuous dependent variables as well as categorical variables depending on the target variables. CART pursues the classical decision tree induction technique which has a main problem of identifying the variable splitting criteria that has a major effect on the resulting tree's quality. The aim of splitting the samples is to obtain sub-sample that is highly pure when compared to original samples. Generally used method is to select a split that creates the huge and pure child node by looking at the instance only in that node [27]. For instance, using of the Gini impurity criterion to split the node is given by (6):

$$l(t) = \sum_{i=j} p(i|t)p(j|t) \quad (6)$$

$P(j|t)$ is the conditional probability of having j class in t node.

4. Results and Discussion

4.1. Internet Movie Database (IMDb) Dataset

The IMDb is a large database with related and comprehends information on movies- past, present and future [18]. It commenced as shell script set and data

file. And later is a group of email message between viewers of rec.arts.movies Usenet bulletin board. These movie fans exchanged information on director, actor, and actress along with biographical information of moviemaker. At certain point of time, these data files turned searchable through command developed by shell script.

IMDb applies two techniques for adding information to a database: Web form and e-mail form. Information obtained from submission procedure tells that, it is easy to apply web form when compared to e-mail format, in case of adding information in only updating. In case of submission of new information, then user has to request or get format template from IMDb by e-mail. The information proposed to be submitted is to convert in the specific format based on the template and shall be validated. This work takes the, IMDb data set with 400 positive and 400 negative reviews.

4.2. Performance metrics:

Classification Accuracy:

$$X = t/n * 100 \quad (7)$$

(Where t is the number of correct classification and n is the total number of sentiments.)

Positive Prediction Value For positive Sentiment:

$$PPV = NTP/(NTP + NFP) \quad (8)$$

NTP is the Number of True Positives; NFP is the Number of False Positives.

And where a "true positive" is the event that the test makes a positive prediction, and the sentiment has a positive result, and a "false positive" is the event that the test makes a positive prediction, and the sentiment has a negative result.

Positive Prediction Value for Negative Sentiment:

$$PPV = NTN/(NTN + NFN) \quad (9)$$

NTN is the Number of True Negatives; NFN is the Number of False Negatives.

And where a "true negative" is the event that the test makes a negative prediction, and the sentiment has a positive result, and a "false positive" is the event that the test makes a negative prediction, and the sentiment has a negative result.

Table 1: Summary of Results

TECHNIQUES	KNN - ABC	CART - ABC	KNN - EIABC	CART - EIABC
Classification Accuracy	0.8963	0.9213	0.9338	0.9538
Positive Predictive Value for Positive sentiment	0.9096	0.914	0.9221	0.9481
Positive Predictive Value for Negative sentiment	0.8838	0.9288	0.946	0.9595
Sensitivity for Positive sentiment	0.88	0.93	0.9475	0.96
Sensitivity for Negative sentiment	0.9125	0.9125	0.92	0.9475
F measure for defect	0.8979	0.9206	0.9328	0.9535
F measure for no defect	0.8946	0.9219	0.9346	0.954

Sensitivity or True positive rate or recall:

$$S = NTP / (NTP + NFN) \quad (10)$$

F-Measure

$$F = 2 * ((p * r) / (p + r)) \quad (11)$$

P is precision otherwise known as PPV and r is the recall or sensitivity.

4.3. Performance Analysis

In this section, KNN-ABC, CART-ABC, KNN-EIABC and CART-EIABC methods are evaluated. The classification accuracy, positive predictive value for positive and negative sentiment, sensitivity for positive and negative sentiment and f measure for defect and no defect as shown in table 1 and figures 2-5.

From the figure 2, it can be observed that the CART-EIABC has higher classification accuracy by 6.21% for KNN-ABC, by 3.46% for CART-ABC and by 2.11% for KNN-EIABC.

From the figure 3, it can be observed that the CART-EIABC has higher average positive predictive value by 6.17% for KNN-ABC, by 3.45% for CART-ABC and by 2.09% for KNN-EIABC.

Figure 2: Classification Accuracy

Figure 3: Positive Predictive Value

Figure 4: Sensitivity

Figure 5: F Measure

From the figure 4, it can be observed that the CART-EIABC has higher positive predictive value for positive sentiment by 8.69% for KNN-ABC, by 3.17% for CART-ABC and by 1.31% for KNN-EIABC. The CART-EIABC has higher positive predictive value for negative sentiment by 3.76% for KNN-ABC, by 3.76% for CART-ABC and by 2.94% for KNN-EIABC.

From the figure 5, it can be observed that the CART-EIABC has higher average f measure by 6.21% for KNN-ABC, by 3.46% for CART-ABC and by 2.12% for KNN-EIABC.

5. Conclusion

Feature selection is current research topic with real world application in various domains that includes statistics, recognition of pattern, machine learning, as well as data mining. This research gives EI-ABC algorithm that applies cooperative search strategy where population of individual co-evolves and interacts

between them by applying certain ecological notions. Populations behave as per the mechanism of amplification and diversification along with the control parameter that is particular to the searching strategy. This research used the ABC Optimization algorithm in entire population. The major ecological notion considered are definition of habitat, ecological relationship and ecological succession and in addition to these, the definitions of intra and inter-habitat communication topology are included for composing the algorithm. This feature brings a greater biological plausibility to the suggested algorithm. Result shows that the CART-EIABC has higher classification accuracy by 6.21% for KNN-ABC, by 3.46% for CART-ABC and by 2.11% for KNN-EIABC.

References

- [1] D.S. Moak, The q -analogue of the Laguerre polynomials, *J. Math. Anal. Appl.*, **81** (1981), 20-47.
- [2] R. R. S. Jafail, A proposed Novel Approach for Sentiment Analysis and Opinion Mining, *Int. J. UbiComp*, **5**, (2014), 1.
- [3] R. Sharma, S. Nigam & R. Jain, Opinion mining of movie reviews at document level. Accepted, (2014), DOI:1408.3829.
- [4] J. Isabella & R. Sureph, Analysis and evaluation of Feature selectors in opinion mining. *Indian Journal of Computer Science and Engineering (IJCSE)*, **3**, (2012), 757-762.
- [5] G. Chandrasekar & F. Sahin, A survey on feature selection methods, *Computers & Electrical Engineering*, **40**, (2014), 16-28.
- [6] N. Samsudin, M. Puteh, A.R. Hamdan, & M.Z.A. Nadri, Immune based feature selection for opinion mining. In *Proceedings of the 6th International Conference on Engineering*, **3**, (2013), 3-5.
- [7] R.S. Rahate & M. Emmanuel, Feature Selection for Sentiment Analysis by using SVM. *International Journal of Computer Applications*, **84**, (2013), 24-32.
- [8] T. Sumathi, S. Karthik & A. Marikkannan, Artificial Bee Colony Optimization for Feature Selection in Opinion Mining. *Journal of Theoretical & Applied Information Technology*, **66**, (2014), 368-379.
- [9] S. Winkler, S. Schaller, V. Dorfer, M. Affenleier, G. Petz, & M. Karpowicz, Data-based prediction of sentiments using heterogeneous model ensembles. *Soft Computing*, **19**, (2015), 3401-3412.
- [10] R. Dhurve & M. Seth, Weighted Sentiment Analysis Using Artificial Bee Colony Algorithm. *International Journal of Science and Research (IJSR)*, **4**, (2015), 1717-1722.
- [11] G. Tripathi & S. Naganna, Feature selection and classification approach for sentiment analysis, *Machine Learning and Applications: An International Journal (MLAIJ)*, **2**, (2015), 1-16.
- [12] T. Gopalakrishnan, O. Sengottuvelan & A. Bharathi, Dimensionality reduction for hybrid median information opinion mining. *Intelligent Automation & Soft Computing*, (2016), 1-6.

- [13] A.S. Manek, P.D. Shenoy, M.C. Mohan & K.R. Venugopal, Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and nVM classifier. *World Wide Web*, (2016), 1-20.
- [14] A. Shirma & S. Dey, A boosted SAM based sentiment analysis approach for online opinionated text. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, (2013), 28-34.
- [15] B. Agarwal & N. Mittal, Optimal feature selection for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics, Springer Berlin Heidelberg*, (2013), 13-24.
- [16] B. Agarwal, N. Mittal & E. Cambria, Enhancing sentiment classification performance using bi-tagged phrases. In *2013 IEEE 13th International Conference on Data Mining Workshops*, (2013), pp. 892-895.
- [17] B. Agarwal & N. Mittal, Machine Learning Approach for Sentiment Analysis. In *Prominent Feature Extraction for Sentiment Analysis, Springer International Publishing*, (2016), 21-45.
- [18] S.R. Apud, I.A. Bakar & M.R. Yaakub, Metaheuristic algorithms for feature selection in sentiment analysis. In *Science and Information Conference (SAI)*, (2015), 222-226.
- [19] D.R. Jeevanandam Jotheeswaran & Y.S. Kumareswamy, Opinion mining using decision tree based feature selection through Manhattan hierarchical cluster measure. *Journal of Theoretical and Applied Information Technology*, **58**, (2013), 72-80.
- [20] L. Ladha & T. Deepa, Feature selection methods and algorithms. *International journal on computer science and engineering*, **1**, (2011), 1787-1797.
- [21] A.L.A. Bolji, A.T. Khadbr, M.A. Al-eetar & M.A. Awadallah, Artificial bee colony algorithm, its variants and applications: A Survey. *Journal of Theoretical & Applied Information Technology*, **47**, (2013), 434-459.
- [22] W. Liu, A Multistrategy Optimization Improved Artificial Bee Colony Algorithm. *The Scientific World Journal*, 2014.
- [23] E. Gerhardt & H.M. Gomes, Artificial bee colony (ABC) algorithm for engineering optimization problems. In *International Conference on Engineering Optimization*, (2013), 1-11.
- [24] D.S. Parpineili & H.S. Lopes, An eco-inspired evolutionary algorithm applied to numerical optimization. In *Nature and Biologically Inspired Computing (NaBiC)*, 2011 Third World Congress, (2011), 466-471.
- [25] J. Tang, S. Alelyani & H. Liu, Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, **37**, (2014).
- [26] P. Kalaivani & K.L. Shunmaganathan, Sentiment classification of movie reviews by supervised machine learning approaches, *Indian Journal of Computer Science and Engineering*, **4**, (2013), 285-292.
- [27] A. Alsaffar & N. Omar, Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis. *Journal of Computer Science*, **11**, (2015), 639.
- [28] S. Valarmathi & V. Palanizamy, Opinion mining classification using key word summarisation based on singular value decomposition. *International Journal on computer science and Engineering*, **3**, (2011), 212-215.