Medical Blog Classification Using Hybrid Feature Selection Mechanisms

Neeba E.A.[1*] and Koteeswaran S.[2]
1. Department of Information Technology, Rajagiri School of Engineering and Technology, Kochi, INDIA
2. Department of Computer Science and Engineering, Vel Tech Dr. RR and Dr. SR Technical University, INDIA
*neeba.ea.it@gmail.com

Abstract
In general, in web posts, contents are categorized as affective as well as informative content. Posts that have informative content can be considered as medically relevant, while posts with the treatment performed and the opinions of the treatment, illness, medication etc. are considered affective posts. The current study concentrates on classifying the medical blogs as affective or informative. Hybridized feature such as Bacterial Foraging Optimization Algorithm (BFOA) is considered with regards to speed and accuracy. Bacterial Foraging Particle Swarm Optimization (BF-PSO) solves the optimization issues and goes through a feature selection phase for reducing problem features prior to classification. Boosting helps in solving classification issues including cancer classification, text classification etc.

Keywords: Medical blogging, Bacterial foraging optimization (BFO), Bacterial foraging Particle Swarm optimization (BFPSO), Feature selection and Boosting.

Introduction

Medical blogs are write-ups in the medical domain which is comparatively a newer addition. Blogs are instantly accessible to anyone with an internet connection and they have an open-access leading to easy accessibility and are decentralized accentuating their diversity. When compared to conventional media like online news websites blogs are unique in two ways: 1) they are personalized as they are handled by single individuals; 2) the linked structures between blogs lead to localized communities. Increased research work is being done including content based analysis as well as the evolution of blog communities which focuses on the characteristics of the blogs respectively. Many new tools are also offered to users for retrieving, organizing and analyzing the information provided as more and more people have begun to write blogs. Blog-search engines as well as blog tagging systems is also on the rise.

Medical blogs differ depending on the writer – doctor, patient, or nurse. They can be either affective or informative. Affective post relies mostly on personal emotions and feelings while informative posts provide information about the disease condition and more generalized and technical details related to the disease. The posts written by patients about their disease conditions, their personal experiences about the disease symptoms and treatment or exchange other health-related information which are more personalized come under affective post.

From a doctor's perspective, the health information, insights into a clinician's day-to-day activity, and problems related to a patient's treatment offered by him is an affective post. In the same way, nurses write about general information regarding books as well as family and regarding their personal experiences with patient[1]. Thus in an affective post the author describes the activities carried out in a day, their thought about treatment, the disease, medication or feeling. He does not discuss more generalized medical content but only links to other websites. While, a medical post is regarded as helpful, if it comprises generic or disease-specific information, news on current health studies or on general transferable experiences regarding treatment of particular disease. Thus it is helpful to differentiate between blogs on the basis of their medical relevance as well as their informative content.

To define informative and affective articles we did a survey among the users who blog usually to ensure the content they prefer to read. According to the survey: Informative article include news similar to that of conventional web news channels, technical descriptions related to programming techniques, general knowledge and the current affairs. Affective article include diary about personal affairs and self-feeling or emotion descriptions.

As blogs are more personalized, they not only comprise information but also personal content such as feelings, opinions, as well as attitude. Currently, the research activities towards emotions stated in natural language texts are being scrutinized under the subjectivity analysis as well as affective computing[2]. A special lexical resource is needed based on textual keyboard for the subjectivity analysis. Affective lexicon is the most important source in detecting emotions in text and several methods to develop these dictionaries are under construction. The association of data with concept enhances OM and permits more affective data.

Feature selection aids in eliminating irrelevant features. By feature selection the quantity of data can be reduced and the computation can be brought down thus improving the performance of the system. Several optimization techniques such as PSO, ACO, BFOA, are used for selecting the subset features. The biometrics such as finger are palm print are given as input and the dominant feature improves the classification accuracy.

Classification is a form of pattern registration where the classifier identifies the category to which the test sample belongs from a set of samples. The feature with optimum similarity is selected and is trained using three machine learning classifiers. For instance, in the boosting proocol, weak classifiers are trained in a hierarchical manner for learning the hardest problems and thus combining many weak classifiers to become a strong learner.

Related Works

Onan and Korukoğlu[3] made a comparison of OM data sets 5 classifiers. The outcomes reveal that the ensemble technique could help in constructing effective OM classification techniques.

Missen et al[4] proposed a model which uses the primary components of the blogosphere to extract opinions from blog posts. Apart from this, opinions prediction as well as multi-dimensional ranking was emphasizing together with the problems which scholars may face when building the suggested model. Finally, the significance of social networking evidence were demonstrated through experiments.

Zhu et al[5] proposed a novel method for opinion summarizations on the basis of sentences selection. Opinions should be in the form of short summary that is easily accessible at the same time with a few informative sentences from the viewers perceptive. Here it is more of a community leader detection problem with the cluster of sentences considered as a community. The detected leaders may be regarded as the sentences with the greatest amount of information, while informativeness should pertain to both the community as well as the document it is a part of. Review information from 6 product domains from Amazon.com are utilized for verifying the efficacy of the technique of opinion summarization.

Social media has become the floor for communicating sentiments via blog posts, microblog posts, tweets, instant messages, new portals etc. Rao et al[6] performed a research on the detection of emotions from both the writer as well as the reader's perspective. For bridging the gap between the social media as well as the readers' emotion, an intermediary layer was introduced to make an effective model. The suggested model will help in classifying social emotion of unlabelled texts as well as for generating a social emotions lexicon. Exhaustive evaluation utilizing real world data validates the efficacy of the suggested model for the two applications.

Methodology

Mayo clinic dataset: The Mayo Clinic is a nonprofit organization dedicated to various researches and diagnosis and treatment of illnesses. The clinical researchers from Mayo Clinic contribute to the comprehension of the disease process as well as translations of lab results to the clinical practices. Almost 600 research level scientists as well as doctoral level physicians are hired here with extra 3400 health care staff as well as students with appointments in research. In 2015, more than 2700 research protocols were overviewed by the Mayo Clinic Review Board and 11000 ongoing human researches thus leading to 7300 research publications as well as review articles in peer reviewed journals. It hosts an extensive website in which huge quantity of health data as well as tools are given. In this website the patient reviews are longer with respect to words (approximately 387 words) as well as sentences (21 sentences in an average) than the posts by nurses as well as physicians which were only around 13 sentences as well as 300 words on an average.

Feature Extraction: Features which occur rarely in a webpage reflect the categories of webpages better. In IR, for identifying the discriminative words, a statistical metric called Term Frequency–Inverse Document Frequency (TF-IDF)[7] is utilized. A calculation of how important is word in a documents set is expressed as term frequency. In a documents set, TF of a word (t_i) is given by equation (1):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

wherein $n_{i,j}$ refers to the quantity of occurrences of the word (t_i) in a text d_j, while the denominator refers to the sum of the quantity of occurrences of every term in the text d_j. The IDF is a metric that assesses the importance of a word in the entire document collection.

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2}$$

wherein | D | refers to the overall quantity of the texts in the archive and $|\{d : t_i \in d\}|$ refers to the quantity of texts wherein a word ti occurs Then:

$$(TF - IDF)_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

when IDF factor is incorporated, the weight of terms which occur often in the set decreases while the weight of the words that appear rarely rises. As a webpage is transformed into a text file after pre-processing, TF-IDF statistical metric is utilized for selecting the initial collection of rare features F1 that hold a great deal of information of the webpage.

Particle Swarm Optimization (PSO): The method of PSOA is initialized with a set of randomly distributed particles designated with random velocity. The particle y in the d-dimensional problem space, cluster in tandem, and in the converge to a global optimum. The motion of particles in the search space is according to the experiences of the agent as well as its neighbours in the swarm (swarm intelligence). Assume the ith particle at the subsequent iteration is $x_{id}(t+1)$

and $V_{id}(t+1)$ respectively, which can be given mathematically as in equation (4):

$$\begin{aligned} &V_{id}(t+1)=w.\mathrm{V}_{id}(t)+\mathrm{c}_1.r_1[p_{id}(t)-x_{id}(t)]+\mathrm{c}_2.r_2[\mathrm{g}_d(t)-x_{id}(t)],\\ &x_{id}(t+1)=x_{id}(t).V_{id}(t+1) \end{aligned} \tag{4}$$

In this formula, variable is called the inertia constant which maintains a balance between local as well as global searches, c_1 and c_2 are acceleration constants. r_1 and r_2 Refer to 2 independent created arbitrary numbers that are uniformly spread between [-1,1]. $p_{id}(t)$ Denotes coordinates of the best position found so far by the ith particle (local optimum), while the coordinates of the best position found so far by the whole swarm (global optimum) are stored in $g_d(t)$.

The exploration of fresh search space relies on the value of inertia constant. Hence, Eberhart and Shi suggested a modified which reduces linearly with successive cycles[8], that may be expressed as in equation (5):

$$w=w_{\max}-(w_{\max}-w_{\min})\frac{g}{G} \tag{5}$$

Here g is the generation index denoting the current quantity of evolutionary generations, G refers to the preset value of maximal quantity of generations, $w_{\max}$ and $w_{\min}$ refers to the maximum as well as minimum weights. Initial value of ω is 0.9 to permit the particles for finding the global optima neighbourhood more rapidly. Value of ω is initialized at 0.4 on discovering the optimum such that the search is moved from exploration to exploitation. Search procedure is ended when there is no more enhancement in the global optima or the quantity of cycles completed is the same as the final pre-set value.

```
"Step 1: Set the BFO parameters,
N: Total Bacteria, $N_C$ : Count of Chemotaxis steps, $N_{re}$ : Total reproductive steps, N: Dimensions of the problem, C: Step size taken in tumbling, ω: The inertia weight, $C_1$ : Swarm Confidence, X(i): Location of the ith bacterium, V(i):Velocity of the ith bacterium, Gbest: Global best position value and Pbest: Local best position value
Step 2: Begin Elimination dispersal loop
Step 3: For every reproduction step do
Step 4: For every chemotaxis step do
    a. Calculate the fitness function (J) of the initial population
    b. Set Jlast=J.Hold this value to find better cost value via a swim
    c. Tumble: Create a arbitrary vector delta -1 to 1.
    d. Move: Let move the bacterium to a position with step size
    C(i) using the below equation called Tumble
    Del=(rand(1,1)-0.5)*2
    $x(i+1)=x(\mathrm{i})+\mathrm{C}(\mathrm{i})\frac{Del(i)}{Del(i)Del^T(i)}$
    e. Again swim
    Assume m=0
    While m< $N_s$
    M=m+1
    If J(i)>Jlast
    Let Jlast=J(i) and use this Jlast to calculate new J(i)
    Assume m= $N_s$
    End
    Step 5: Mutation (by PSO operator)
    For i=1,2,….,S
    Initialize the Gbest and pbst
    Update the position as well as velocity of the ith bacterium using to the following equations:
    $V_i(i+1)=\omega V_i(i)+C_1\phi_1(pbest-X(i))+C_2\phi_2(\mathrm{g}best-X(i))$
    $\mathrm{X}(\mathrm{i}+1)=\mathrm{X}(\mathrm{i})+\mathrm{V}_i(i+1)$
    Step 6: Let $S_r=S/2$
    The bacteria with the lowest Jhealth (final fitness) values will die and the remaining bacteria with the best fitness values are split in to two bacteria thus making population of bacteria constant."
```

Figure 1: Pseudo code for Hybrid Particle Swarm Optimization (PSO)-Bacterial Foraging Algorithm (BFOA)

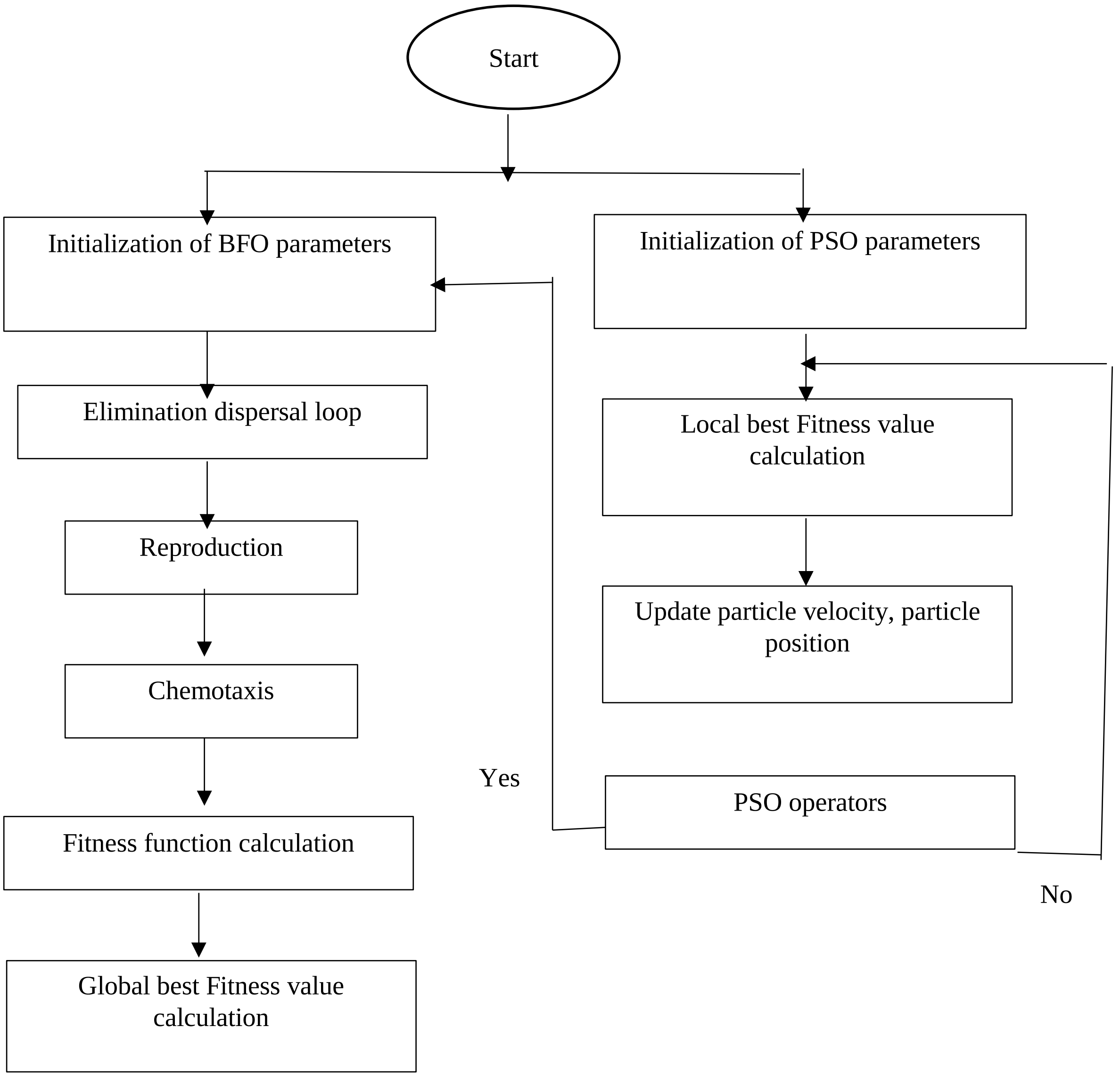
Start
Initialization of BFO parameters
Elimination dispersal loop
Reproduction
Chemotaxis
Fitness function calculation
Global best Fitness value calculation
Initialization of PSO parameters
Local best Fitness value calculation
Update particle velocity, particle position
Yes
PSO operators
No

Figure 2: Flowchart for Hybrid Particle Swarm Optimization (PSO)-Bacterial Foraging Algorithm (BFOA) running in parallel

```
For k = 1, 2, …, K Do
Take N(B) instances arbitrarily as well as with substitution from the X.
And create the learning set L(B).
Train classification k (L(B)) from the learning set L(B).
Make a plurality vote for each k of K classifications on the test set.
The classification, that obtains highest voting score, is the best.
```

Figure 3: Bagging Algorithm

Evolutionary Bacterial Foraging Algorithm (EBFO): An evolutionary protocol is based on nature based on the foraging activity of the animals or birds. Natural selection system based on the elimination of animals with poor foraging strategies and the acceptance of species with excellent foraging scheme is propagated. This was studied by Bremermann as well as researched further by Passino for constructing the Bacteria Foraging Optimization Algorithm, based on the bacteria's ability to search for foods with more nutritional levels and discarding the toxic elements using optimum energy may be comprehended as a procedure of optimization. The fundamental operations of EBFO[9] protocol is detailed here:

Chemotaxis: When foraging, E. coli bacteria moves toward food locations through swims as well as tumbles through usage of flagella. The operations are carried out throughout the life cycle.

Swarming: After successfully determining the direction of best food location, bacteria that possesses knowledge of the optimal route to food source attempts to tell this to the other utilizing attraction signals. Bacteria form swarms in the positive nutrients gradient, increasing the bacterial concentrations.

Elimination-Dispersal. On the basis of environment conditions like changes in temperature, toxic settings, as well as accessibility of food, population of bacteria either changes gradually or suddenly. These processes are iterated till optimized solutions are attained.

Boosting utilizing the log-likelihood loss for binary classification and typically for multi-class issues is called LogitBoost. .The LogitBoost protocol comprises the steps given below[11]:

1. Begin with weights $w_i = 1/N, i = 1, 2, ..., N,$ $\text{F(x)=0 and probability estimates p(x}_i) = 1/2$
2. Iterate for m=1,2,….,M
   a. Compute the working response as well as weights:
   $$z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1-p(x_i))}$$
   $$w_i = p(x_i)(1-p(x_i))$$
   b. Fit the function $f_m(x)$ by a weighted least- squares regression of z_i to x_i utilizing weights w_i .
   $$F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$$
   c. Update: $p(x) \leftarrow \dfrac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$
3. Output the classifier:
   $$sign[F(x)] = sign\left[\sum_{m=1}^{M} f_m(x)\right]$$

   Here, sign[F(x)] refers to a function which has 2 possible output labels:
   $$sign\left[F(x)\right] = \begin{cases} 1 & \text{if F(x)<0} \\ -1 & \text{if F(x)} \geq 0 \end{cases}$$

Particle Swarm Optimization-Bacterial Foraging Algorithm (PSO-BFO): A new protocol combing BFO and PSO is proposed which is termed as the BFPSO algorithm[10] with high convergence speed and commendable accuracy. The positive attribute of both BFO and PSO combine to give a better protocol; that is PSO helps in conducting global searches yielding a near optimum solution while BFO fine tunes the solution and gives an option high accuracy. The convergence speed of PSO is high but it has an inherent disability of getting forced into the local optimum where as in BFO the convergence speed is low but does not get trapped in the local optima.

Bagging and Logit Boost: Bootstrap aggregating otherwise commonly called as bagging is an ensemble method for improving the classification schemes. Breiman worked in improving the bagging by using it as a variance reduction technique for a given base procedure, thus fitting in a linear model. This has proved to be successful due to its implementation simplicity and popularity. The main disadvantage of bagging is the lack of interpretation. Figure 3 shows the bagging protocol.

The building of weak classifier is a key factor that affects performance of boosting protocol. Weak classifier $f_m(x)$ in step 2 ought to be capable of coping with re-weighing the data as well as robust against overfitting. Decision trees are adequate for weak classifiers for LogitBoost.

Results and Discussion

For experiments, EBFO-Bagging, BFO-PSO-Bagging, EBFO-Logiboost and BFO-PSO Logiboost were compared each other for obtaining classification accuracy, average precision and recall respectively. Table 1 and figure 4 to 6 shows the same.

From table 1 and figure 4 it is show that the classification accuracy of BFO-PSO Logiboost performs better by 2.95% than EBFO-Bagging, by 1.81% than BFO-PSO-Bagging and by 2.26% than EBFO-Logitboost.

From table 1 and figure 5 it is show that the average precision of BFO-PSO Logiboost performs better by 3.08% than EBFO-Bagging, by 1.67% than BFO-PSO-Bagging and by 2.49% than EBFO-Logitboost.

Table 1
Summary of Results

		BFO-PSO	EBFO-	BFO-PSO
	EBFO- Bagging	**Bagging**	**Logitboost**	**Logitboost**
Classification accuracy	0.9158	0.9263	0.9221	0.9432
Average				
Precision	0.9084	0.9213	0.9138	0.9368
Average Recall	0.9099	0.923	0.9149	0.9402

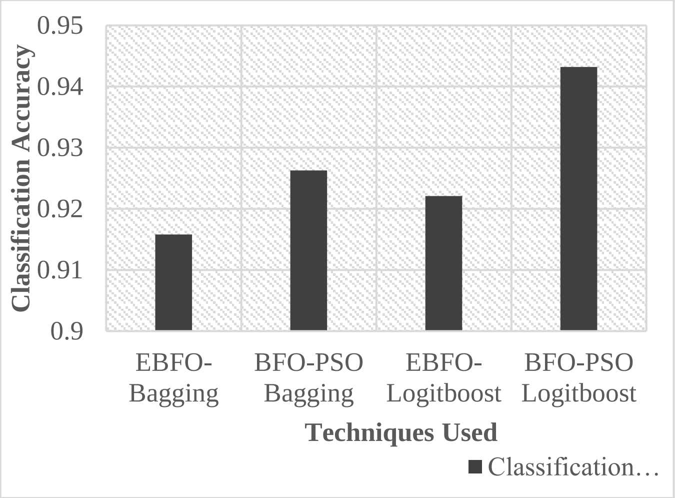

Figure 4: Classification accuracy

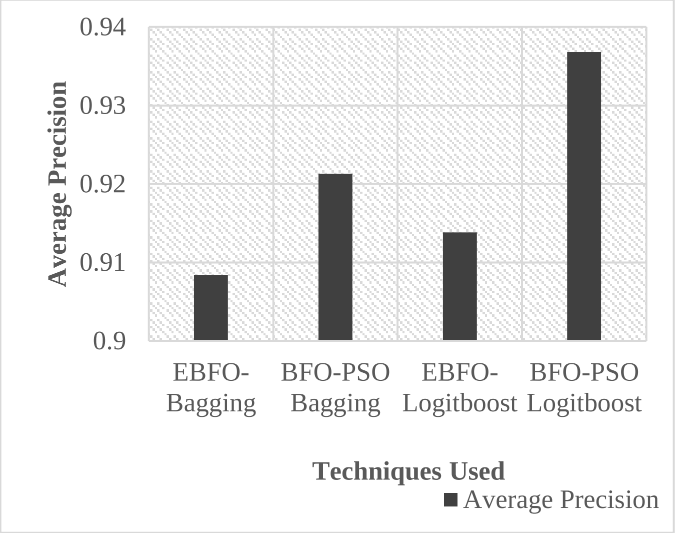

Figure 5: Average Precision

From table 1 and figure 6 it is show that the average recall of BFO-PSO Logiboost performs better by 3.28% than EBFO-Bagging, by 1.85% than BFO-PSO-Bagging and by 2.73% than EBFO-Logitboost.

Conclusion

In this work, feature selection techniques were compared with classification techniques such as bagging and Logitboost. Selecting the best features for building the classifier is more significant than modeling the classifier itself. LogitBoost reduces the training errors linearly and hence yield better generalization.

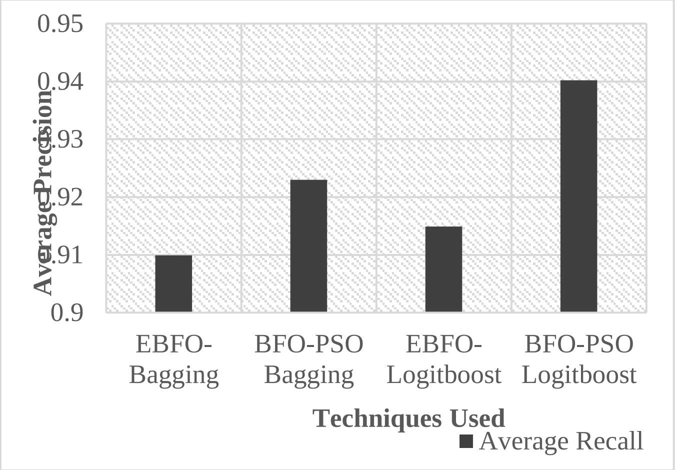

Figure 6 Average Recall

References

1. Das D., Studies on emotion analysis at word and sentence level, Lambert Academic Publishing, 96, **(2011)**

2. Poria S., Gelbukh A., Cambria E., Hussain A. and Huang G.B., Emo Sentic Space: A novel framework for affective common-sense reasoning, *Knowledge-Based Systems*, **69**, 108-123 **(2014)**

3. Onan A. and Korukoğlu S., Ensemble methods for opinion mining, In 2015 23nd Signal Processing and Communications Applications Conference (SIU), IEEE 212-215, **(2015)**

4. Missen M.M.S., Boughanem M. and Cabanac G., Opinion detection in blogs: what is still missing?, In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference, 270-275 **(2010)**

5. Zhu L., Gao S., Pan S.J., Li H., Deng D. and Shahabi C., Graph-based informative-sentence selection for opinion summarization, In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference, IEEE, 408-412 **(2013)**

6. Rao Y., Li Q., Wenyin L., Wu Q. and Quan X., Affective topic model for social emotion detection, *Neural Networks*, **58**, 29-37 **(2014)**

7. Mangai J.A. and Kumar V.S., A novel approach for web page classification using optimum, *IJCSNS*, **11(5)**, 252 **(2011)**

8. Patnaik S.S. and Panda A.K., Particle swarm optimization and bacterial foraging optimization techniques for optimal current harmonic mitigation by employing active power filter, *Applied Computational Intelligence and Soft Computing*, **2012**, 1 **(2012)**

9. Das S., Biswas A., Dasgupta S. and Abraham A., Bacterial foraging optimization algorithm: Theoretical foundations, analysis, and applications, *Foundations of Computational Intelligence*, **3**, 23–55 **(2009b)**

10. Chen C.H., Su M.T., Lin C.J. and Lin C.T., Hybrid of bacterial foraging optimization and particle swarm optimization for evolutionary neural fuzzy classifier, *Int. J. Fuzzy Syst*, **16(3)**, 422-433 **(2014)**

11. Friedman J., Hastie T. and Tibshirani R., Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics*, **28(2)**, 337-407 **(2000)**.