IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 13, 2019, accepted June 23, 2019, date of publication July 1, 2019, date of current version July 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925905

An Adaptive Model for Identification of Influential Bloggers Based on Case-Based Reasoning Using Random Forest

YOUSRA ASIM[1], BASIT RAZA[1], AHMAD KAMRAN MALIK[1], AHMAD R. SHAHAID[1], AND HANI ALQUHAYZ[2]

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad 45550, Pakistan
[2]Department of Computer Science and Information, College of Science in Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia

Corresponding author: Basit Raza (basit.raza@comsats.edu.pk)

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project Number No. 1440-135.

ABSTRACT Bloggers play a role in individual decision making of users in online social networking platforms. Their capability of addressing a wide audience gives them influence over their audience, which companies seek to exploit. Identification of influential bloggers can be seen as a machine learning (ML) task and different ML techniques can help in classifying the professional blogger. In this paper, we propose a predictive and adaptive model named as Influential Blogger based Case-Based Reasoning (IB-CBR) model for the recognition of unseen influential bloggers. It incorporates self-prediction and self-adaptation (self-management) capabilities which are the essence of an automated system. The integration of Random Forest is found contributing to the efficiency of the IB-CBR model as compared to Nearest-Neighbor, and Artificial Neural Network. The performance of the proposed IB-CBR model is evaluated against other ML techniques by using standard performance measures on a standard blogger's dataset. It is observed that our proposed model exhibits 88–95% Accuracy and 94–97% True Positive Rate in the prediction and adaptation of professional bloggers, respectively, in three iterations of the proposed model. What's more, the IB-CBR model achieved 91–96% (increasing) F-measure, 91–98% (increasing) ROC AUC, and 36–11% (decreasing) False Positive Rate due to adaptivity. The IB-CBR model performed well when it is compared with other ML techniques using different standard datasets.

INDEX TERMS Blogging, blogger classification, case based reasoning (CBR), machine learning.

I. INTRODUCTION

Online Social Network (OSN) is a universal platform for the people of all races, classes, and nationalities to show their views and experiences. It can bring people together living far and wide. Individuals can share their views, follow trends, like or unlike ideas. Blogging is one of the well-known OSN services through which bloggers not only share their ideas and opinions by writing blogs, but also build strong bonds with their followers. The visitors of their blogs can interactively participate online by reading and leaving positive or negative comments on their blogs. Not all bloggers are equally supported by their blog readers. A few bloggers are appreciated the most by their addressees as compared to their rival bloggers. The individuals who attract readers by their thoughts, ideas, suggestions, and opinions are considered influential [1]. Although, the influential nodes are smaller in number, they are effective in influencing others and controlling the social network [2], [3]. Such nodes have been used in social networks for viral marketing [4], targeted marketing [5], propagation of brand information [6], [7], brand advertisement and purchases [8]–[10], online campaigns [11], influence maximization [12]–[14], and information diffusion [15]–[18].

Specifically speaking, in the blog network, Booth and Matic [10] has contended that ultimately the bloggers with a wide audience use their power to affect the thoughts and feelings for advertising particular brands or projecting

The associate editor coordinating the review of this manuscript and approving it for publication was Bilal Alatas.

certain ideas. They have an ability to exert influence on customers and can be a productive source in any social media campaign. Marketing companies are often on the lookout for such influential bloggers as they can be of immense help in marketing their brands. It is suggested that the bloggers can be differentiated among professionals and amateurs depending upon their blogging contributions [19]. The professional bloggers have an in depth, thorough approach towards one or the other areas of interest. They have a deeper understanding of the mass psychology with an adequate experience level. On the other hand, amateur bloggers have a partial knowledge towards a random topic range and their unsystematic way of blogging fails to achieve the interest of the more serious audience. Naturally the marketing companies aim at engaging a professional blogger who is more dedicated towards blogging. But the process of identifying professional vs non-professional is not so easy.

In the past, network based and feature based models have been suggested for influential blogger's identification [20], [21], and also by using a few Machine Learning (ML) techniques for the classification of influential bloggers based on the labeled data [22]–[24]. This research focuses on the latter studies only. Though, authors have suggested the use of ML techniques in the context of influential bloggers identification as a future direction due to their little use [25] however, it has been noticed that previously used ML techniques for blogger classification [22]–[24] were not adaptive for automatic identification of unseen influential bloggers based on varying factors. What's more, it only applied different available ML techniques instead of proposing a new intelligent algorithm for the classification of bloggers into professional or otherwise.

Being an extension to our prior work, the objective of this study is to propose a new prediction algorithm for the identification of influential bloggers based on their features. Besides, another objective is to add adaptive capabilities in the proposed algorithm to adapt the changes in the changing behavior of blogger which is found as a limitation in the previous studies. For this purpose, we have focused on autonomic characteristics which can enable a system to work independently by reducing human interference and to manage adaptation behavior [26]. Autonomic systems follow natural phenomenon in the sense that humans go through unknown scenarios in a dynamic environment throughout their life. In such a situation, they have to be in a continuous learning process. Similarly, autonomic systems are enforced to learn by using their self-management capabilities upon facing new experiences. That's why, in such a case, learning is kept dynamic instead of static to cope with undiscovered circumstances. Furthermore, Kephart and Chess [27] have discussed that automonous systems become self-managing having properties like self-protection, self-configuration, self-healing. On the other hand, most of the machine learning techniques build a model without adaptation to devise a solution in future cases which make them infeasible to fulfill the characteristics of autonomic systems. A number of studies have been performed which opt for Case-based Reasoning (CBR) for enjoying the benefits of self-management capabilities [26].

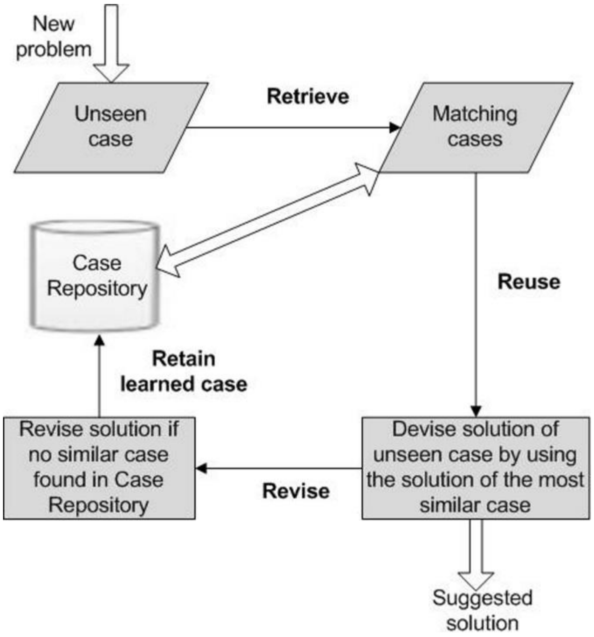

New problem
Unseen case
Retrieve
Matching cases
Case Repository
Reuse
Retain learned case
Revise solution if no similar case found in Case Repository
Revise
Devise solution of unseen case by using the solution of the most similar case
Suggested solution

FIGURE 1. CBR Cycle and its Phases diagram [31].

With all this in mind, we aim at proposing an intelligent classifier based on CBR which is considered as a reasonable technique in the context of adaptation. It can suggest the solution to unseen problems based on the past experience of problem solving. Due to its flexible case structure and proposed solutions, it can perform well in the absence of well-structured knowledge. It can revise the solution according to the newly arrived problem and this revised solution is adapted subsequently. It facilitates to maintain past experiences in problem-solution pair in case-base for future use. Since, it is seen that a number of factors may affect the blogging behavior of a blogger [28]–[30], using CBR in a dynamic manner for the identification of influential bloggers is unique and is suitable for identification of influential blogger. To the best of our knowledge, use of CBR for the identification of influential bloggers based on their features in a dynamic fashion is novel; hence there is a need to investigate how CBR can be used in this problem domain.

There are four phases of CBR namely: retrieve, reuse, revise/adapt and retain as shown in FIGURE 1. The knowledge base contains previous cases. When a new case comes in, it is matched with the existing knowledge base. The case is retrieved from the knowledge base if the match is determined and the suggested solution is reused. Otherwise, the solution for new problem needs revision for better adaptation and a new solution is offered. This revised solution (learned case) is retained in the knowledge base for future purposes.

This study theoretically contributes by suggesting an adaptive model named IB-CBR for influential blogger identification which can accurately predict influential blogger along with its adaptive capabilities for future predictions. The properties of autonomic systems used in IB-CBR are self-prediction, self-adaptation, and self-management. Having prediction abilities, IB-CBR is capable of self-adaptation which includes the adaptation according to the varying patterns in the blogger features. The proposed model is good enough for analyzing new problems and to offer advice in the absence of human intervention which shows its self-management capability. Standard performance measures are used for evaluation and adaptation assessment. The results of the proposed model are compared with different machine learning techniques. Different standard datasets have been used to evaluate the efficiency of IB-CBR for prediction.

The organization of the paper is as follows. Section II presents previous work in the context of the identification of influential bloggers. Section III offers a complete methodology, including the brief note on CBR, dataset description, the algorithms used for results comparison with IB-CBR, the evaluation metrics used for results comparison, and similarity measures used for finding similarity. Section IV presents the proposed IB-CBR model for identification of influential blogger. Section V provides results and discussion. Finally, the Section VI wraps up this study.

II. RELATED WORK

This section provides a glimpse of relevant work for the identification of influential bloggers. The available models have been categorized into Network-based and Feature-based models [20], [21]. Network-based models use several network centrality measures such as degree centrality and closeness centrality to find the influence of a blogger by analyzing the network connections (links) between users. On the other hand, Feature based models emphasize on the characteristics of blog posts of a blogger such as number of comments and blog post length to find his/her influence. Furthermore, these models can be temporal or non-temporal on the basis of the calculation factors used for finding influence of a blogger. Temporal studies are based on finding recentness of the under investigation blog post characteristics, however, non-temporal studies neglect the recency of such features. Recently, Awotunde and Jimoh [32] have proposed a model for the identification of influential bloggers by using social proof, mining their comments and by focusing on their topic of interests.

Aside from that, a few studies exist which use different at hand ML techniques for classification of influential bloggers due to the availability of labeled data by using standard BLOGGER dataset. For instance, the study [23] applied C4.5 decision tree algorithm by using Weka tool and found 82% accuracy for classification task. Likewise, 88% accuracy is observed in case of using Random Forest classifier [24]. Also, K-Nearest Neighbor classifier and Artificial Neural Network have achieved 84% and 90% accuracy respectively [22]. But, in all these studies, results are not validated by performing k-fold cross validation for blogger classification, which is a standard practice in research community to evaluate the performance of a classifier. None of these studies have suggested new algorithms in this context. Asim et.al [30] was an initial effort to overcome the former deficiency and they have investigated different decision tree algorithms, ensemble learning algorithms, and lazy learning algorithms with proper result validation. It was found that Random Forest and Nearest Neighbor achieved 85% accuracy for blogger classification. Besides, 2% gain in accuracy is seen by using Artificial Neural Network [33] coupled with cross validation of results in a standard way. Besides, some of context specific studies are also available in this domain. As an illustration, influential bloggers are discovered by collecting data from Spanish fashion bloggers through questionnaire. It is found that influential bloggers are habitual to reading fashion magazines, they keep on updating their blogs, and work together with media about fashion and fashion events. Such bloggers are active on the web and in their social circles which enable them to be influential in online as well as offline fashion environment [34]. A framework is also proposed to facilitate companies in information diffusion by exploring the influence of a blogger. For this purpose, blog contents and blogger's information both are used to evaluate their influence on Weibo platform to validate the framework [35]. However, the need of an adaptive and a new classifier for influential blogger classification is still there.

III. RESEARCH METHODOLOGY

In this study, we propose an adaptive model that can adapt to new changes in trends among the bloggers and may help in the classification of hitherto unseen bloggers, and also in ascertaining their influence. It is discussed earlier, that CBR methodology will be used which gains experience based on reasoning and it can provide adaptivity in the sense that if a new case is seen during prediction then this approach updates its rules for future cases [36]. In this study, a standard BLOGGER[1] dataset collected [23] is used for experiments. Different results of the state-of-the art ML techniques from previous studies have been used for comparison with IB-CBR model outcomes based on the standard evaluation criteria of Accuracy, TP Rate, FP Rate and F-measure. A comparative analysis of IB-CBR model with adaptation and without adaptation is also performed to observe the performance. A number of well-known similarity measures such as Jaccard, Cosine, Euclidean, Braycurtis, and Canberra distance are used to find the similarity between previously seen scenarios and new problems and to determine the most suitable similarity measure in the selected problem domain. The efficiency of IB-CBR model is examined on different online datasets.

A. DATASET DETAILS

In this study, we used the standard aforementioned Bloggers dataset which consists of five input features namely Degree, Political Caprice, Topics, Local Media Turnover (LMT), and Local, Political, and Social Space (LPSS) and a binary output

[1] https://archive.ics.uci.edu/ml/datasets/BLOGGER

class variable namely Professional blogger, which is either true or false. Without any missing values, this dataset contains 100 instances, including 68 positive instances and 32 negative instances. The details of above-mentioned attributes are as follows:

Degree represents the education level of a blogger. 'Low' value of degree shows that blogger is less educated, 'Medium' value of degree indicates that blogger has B.Sc. level education, and 'High' value of degree indicates that the blogger has M.Sc., and/or Ph.D. level education.

Political Caprice shows the political affiliation of a blogger. 'Left' value of political caprice shows that a blogger is affiliated with reformist party, 'Right' value of political caprice shows that a blogger is affiliated with the conservative party, and 'Middle' value shows no political interests of a blogger.

Topics indicate the area of interest of a blogger with respect to his/her blogging. 'Impression' value of the topic indicates that a blogger is involved in writing his personal experiences in his blogs, 'Political' value shows that a blogger is interested in writing political blogs, 'Tourism' value shows that a blogger is keen towards writing his travelling experiences, 'Scientific' value shows that a blogger is concerned with technical blog writing, 'News' value shows that a blogger is focused on daily updates in his blogs.

Local Media Turnover (LMT) represents whether a blogger rely on the effect of local media on blog writing or not by having two possible values; 'Yes' or 'No'.

Local, Political, and Social Space (LPSS) represents whether a blogger keeps the faith in the effect of local, political, and social conditions on blogging or not by having two possible values; 'Yes' or 'No'.

Professional Blogger is a target output class having two possible values; 'Yes' or 'No', where Yes shows a blogger is professional and No shows that a blogger is non-professional.

B. ALGORITHMS USED FOR COMPARISON

In this work, we have used the following ML algorithms which have been used previously [23], [33] for the classification of bloggers into professional or otherwise on the same dataset. The CBR outcomes are compared with the results of these algorithms. The algorithms which were used in this comparative analysis are as follows:

1) RANDOM FOREST CLASSIFIER

Random Forest (RF) classifier is an ensemble classifier which performs classification by generating a number of decision trees. For this purpose, it randomly selects the number of attributes to produce forest of decision trees which increases its strength over single classifiers for data classification [37]. RF classifier can pick up the best hypothesis when it is applied to a small dataset due to manipulation of several initial points for identifying an unknown function. Due to this fact, we have selected this classifier for blogger classification.

2) NEAREST NEIGHBOR CLASSIFIER

The Nearest-Neighbor classifier (IB1) is a well-known lazy learning algorithm which predicts the output class of a new testing instance by finding its closest previous training instances by using Euclidean distance. Afterwards, the output class of testing instance is suggested based on the class of the closest training instance [38]. In the case of more than one closest training instances, the class of first one is used to predict the class of testing instance. If there exist relevant attributes of both types of instances, then this classifier can perform outstandingly.

3) ARTIFICIAL NEURAL NETWORK ALGORITHM

The Artificial Neural Network (ANN) is considered an efficient tool for classification. It works on the principle of human brain, which is the network of neurons. It consists of interconnected neurons/ perceptrons which make it feasible in the case of non-linearly separable problems. These neurons are organized into three types of layers named as an input layer, hidden layer (can be more than one) and an output layer. The connection between neurons are assigned with different weights [39]. The principle of backpropagation is used to back propagate the error in a sort of credit assignment task, which in turn uses gradient descent to minimize the squared error between network predicted output and the target output. The weights assigned to different connections are kept on being updated until the aforesaid error is minimized. This capability of ANN makes it adaptive and it's the basic reason of using it in this work.

4) C4.5 ALGORITHM

C4.5 algorithm is considered as a statistical classifier which is the extension of ID3 algorithm [40]. It produces decision tree based on training instances (already classifier samples) to predict future instances. It uses information gain of attributes to split each node for producing decision tree. The attribute with the highest information gain value is used to make a decision. It is the most widely used algorithm by practitioners. In this work, we shall compare the results of Gharehchopogh and Khaze [23] for the same dataset of bloggers with CBR outcomes.

C. EVALUATION METRICS

In this work, we have used standard performance metrics such as Accuracy, True Positive Rate (TP Rate), False Positive Rate (FP Rate), F-measure, and ROC area under the curve (AUC) to evaluate the effectiveness of the proposed CBR model and to compare its results against the outcomes provided by different algorithms.

Accuracy indicates the number of correct predictions made by the model over all kinds of predictions made. Equation (1) can be used to calculate accuracy where 'TP' denotes True Positives, 'TN' means True Negatives, 'FP' indicates False Positives, and 'FN' shows False Negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

TABLE 1. Similarity measures used for the calculation of similarity between cases.

Name of Similarity Measure	Definition
Jaccard similarity	$Sim_{ab} = \left\lvert \frac{X_a \cap X_b}{X_a \cup X_b} \right\rvert$ Where Sim_{ab} denotes similarity between a^{th} and b^{th} cases with respect to all features of blogger feature vector.
Cosine similarity	$Sim_{ab} = \frac{\sum_{k=1}^{n} x_{ak} x_{bk}}{\sqrt{\sum_{k=1}^{n} (x_{ak})^2 \sum_{r=1}^{n} (x_{br})^2}}$ Where Sim_{ab} denotes similarity between a^{th} and b^{th} cases with respect to all features of blogger feature vector.
Euclidean similarity	$dis_{ab} = \sqrt{\sum_{a=1}^{n} (x_a - x_b)^2}$ Where dis_{ab} denotes distance between a^{th} and b^{th} cases with respect to all features of blogger feature vector.
Braycurtis similarity	$dis_{ab} = \frac{\sum_{k=1}^{n} (x_{ak} - x_{bk})}{\sum_{k=1}^{n} (x_{ak} + x_{bk})}$ Where dis_{ab} denotes distance between a^{th} and b^{th} cases with respect to all features of blogger feature vector.
Canberra similarity	$dis_{ab} = \sum_{k=1}^{n} \frac{(x_{ak} - x_{bk})}{(x_{ak} + x_{bk})}$ Where dis_{ab} denotes distance between a^{th} and b^{th} cases with respect to all features of blogger feature vector.

TP Rate denotes the fraction of bloggers that are actually professional were predicted professional. Equation (2) is used to calculate TP Rate where 'TPR' shows True Positive Rate, and 'AP' means True Positives.

$$TPR = \frac{TP}{AP} \tag{2}$$

FP Rate denotes the fraction of bloggers that are actually non-professional, but found to be professional. Equation (3) is used to find out FP Rate where 'FPR' represents False Positive Rate and 'AN' shows Actual Negatives.

$$FPR = \frac{FP}{AN} \tag{3}$$

F-measure which is the combination of Precision and Recall will ensure that in classifying instances, each class contains points of only one class, e.g. each class has exactly professional or non-professional bloggers, where Precision will tell us what proportion of bloggers that model diagnosed as professional, are actually professional (How many did we catch?). Recall enables us to ascertain the proportion of bloggers that were actually professional and were diagnosed by the model as a professional (How many did we miss?). Equation (4) is used to determine the F-measure.

$$F - measure = 2 * \left(\frac{\Pr ecision * Re\,call}{\Pr ecision + Re\,call} \right) \tag{4}$$

ROC AUC represents capability of the classification model to distinguish between classes. Its ranges from 0 (0%) to 1 (100%). Higher value of AUC represents that the model is better in distinguishing between professional and non-professional blogger and vice versa. Similarly, the ROC AUC values such as 1 (100%), 0.9 (90%), 0.8 (80%), 0.7 (70%), 0.6 (60%), and 0.5 (50%) indicates *perfect, excellent, good, mediocre, poor, and random* classification respectively.

D. SIMILARITY MEASURES

In this paper, six commonly used distance as well as similarity measures such as Jaccard, Cosine, Euclidean, Braycurtis, and Canberra are applied to calculate the similarity between the new instance and the training data in Algorithm 3. The comparison of these measures is carried out on the basis of performance measures discussed in the section III C. The definitions of these similarity functions are provided in Table 1.

IV. PROPOSED IB-CBR MODEL

This section describes the proposed IB-CBR model for identification of bloggers. By briefly discussing the introduction of CBR and its phases, this section suggests algorithms for CBR phases which can identify influential bloggers using tagged data. CBR is a methodology which can suggest the solution to unseen problems based on the past experience of problem solving in a classic way. It is a lazy learning approach.

Salem and Shmelova [41] presented the models based on CBR and provided the way decision making ought to be done and to better recognize what is important in a new situation. Also, in CBR, implementation is reduced to identifying

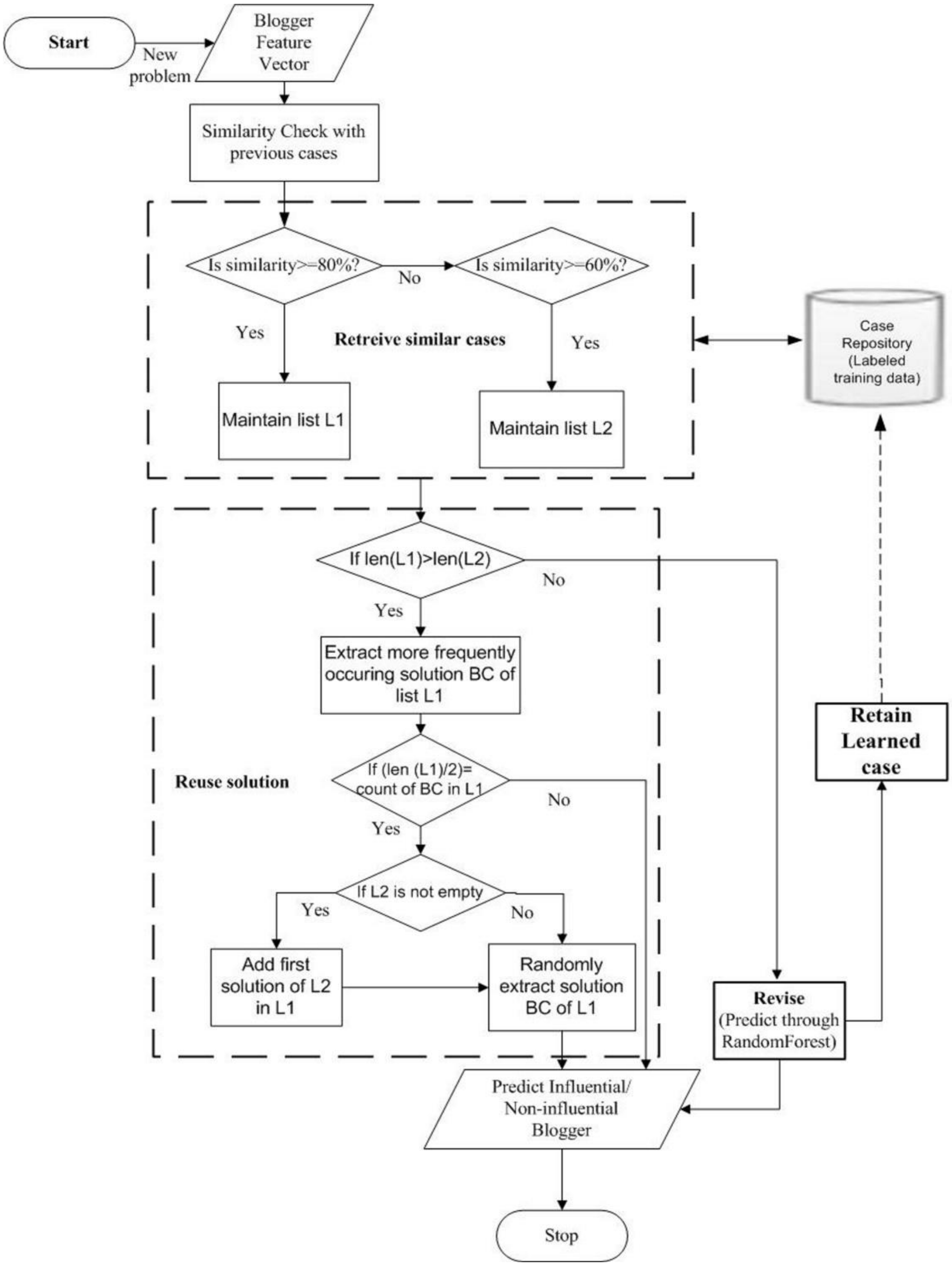

Start
New problem
Blogger Feature Vector
Similarity Check with previous cases
Is similarity>=80%?
No
Is similarity>=60%?
Yes
Retreive similar cases
Yes
Maintain list L1
Maintain list L2
Case Repository (Labeled training data)
If len(L1)>len(L2)
No
Yes
Extract more frequently occuring solution BC of list L1
Retain Learned case
Reuse solution
If (len (L1)/2)= count of BC in L1
No
Yes
If L2 is not empty
Yes
No
Add first solution of L2 in L1
Randomly extract solution BC of L1
Revise (Predict through RandomForest)
Predict Influential/ Non-influential Blogger
Stop

FIGURE 2. IB-CBR model for influential blogger identification.

significant features that describe a case, which is an easier task than creating an explicit model. It is adaptive in the sense that if a new case is seen during the prediction phase then it updates its rules for future cases [36]. For this purpose, problem-solution pair is kept in the case repository which is also known as the case-base. CBR has been used in a number of application domains such as in ecommerce for product selection, help desk applications, medical diagnosis, and software quality prediction, and software reuse [26]. Moreover, in the context of online social networks, CBR has been used for identifying knowledge leaders of particular domains by using their profiles [42].

FIGURE 2 shows the proposed IB-CBR model which consists of four phases namely, Retrieve, Reuse, Revise, Retain, and also a Case Repository. It takes input as Blogger feature vector (new problem) and checks its similarity with existing cases stored in the case-base for the prediction of its solution. When similarity is found equal to or greater than the predefined threshold ($>= 80\%$) and ($<= 80\%$ and $>= 60\%$), similar cases are retrieved and used for prediction depending upon particular conditions, otherwise revise is performed. The new problem and its predicted solution both are retained for future used in the Case Repository.

We have developed a number of algorithms for four CBR phases. The details of each phase and their respective algorithms are presented that gives insights into the proposed model. We used Algorithm 1 for blogger classification in a non-adaptive manner. When a new problem i.e. Blogger feature vector (Bfv) comes for prediction, its similarity is measured with all previous cases (training data) stored in Case Repository (CR).

Algorithm 1 Pseudocode for IB-CBR (Without Adaptation)

```
Input: Given a case repository 'CR' having 'n' cases, blogger feature vector 'Bfv'
Output: blogger class BC (Yes/No)
Method:
for all training cases tc in CR
    if similarity (tc, Bfv) >= 0.80
        maintain list L1 //List of output classes of all matching cases
    end if
end for
BC ← Most_common (L1) // majority class solution is used for prediction
return BC
```

The solutions of all similar cases which have more than 80% similarity with the new problem are stored in list L1. Once all cases are traversed for similarity, then the most frequently occurring solution is determined by Algorithm 5 (see section IV.B) and further used for devising the solution of a new problem.

Likewise, Algorithm 2 is developed for the identification of influential bloggers in an adaptive manner. The number of cases which are kept in case repository (CR), and new problem is taken as input to the IB-CBR algorithm. Initially, two lists namely L1 and L2 are kept for adaptation algorithm instead of one (as maintained in Algorithm 1) and assigned initial values as Null.

The list L1 will be used to store the solutions (output classes) of previously seen problems which have 80% or above similarity with the new problem. The lists L2 will be used to store the solutions of previously seen problems which have 60% and less than 80% similarity with the new problem. RF model is trained (the reason is discussed in section IV. C) based on the available cases in CR for later use. For this purpose, the cases in CR are given as training examples to RF model, each represented by Degree, Political Caprice, Topic, LMT, LPSS, and output class PB. RF generates different decision trees by using different subsets of these input features. We used the default values of decision parameters for RF such as max_depth = 'None' (represents the maximum depth of the tree where the nodes are expanded until all leaves are pure), min_samples_split = '2' (denotes that minimum two number of samples are needed to split an internal node), min_weight_fraction_leaf = 0 (indicates that all the samples have equal weight), max_leaf_nodes = 'None' (indicates that decision trees are grown with unlimited number of leaf nodes), quality of split = 'gini' and 'entropy' (where 'gini' denotes 'Gini gain', and 'entropy' denotes information gain to measure the quality of split), and max_features = 'auto' (represents the number of features used for the best split where auto means max_features = sqrt(n_features)).

If the list L1 returned by Retrieve phase of IB-CBR are not empty, we Reuse the solution (Algorithm 4 is used) otherwise we Revise the solution (Algorithm 6 is used).

There is an idea behind using similarity threshold i.e. 80% for reusing previously solved problem's solution in Algorithm 1 and Algorithm 3. As there are five input attributes in the dataset we are using in this work, if four or five attributes of the new problem are matched with one or more cases of CR, then the solution of a new problem can be predicted on the basis of those previously seen cases. On the other hand, revise is called because reusing the solution in this case is likely to affect the performance of the classifier.

Algorithm 2 Pseudocode for IB-CBR (With Adaptation)

```
Input: Given a case repository 'CR' having 'n' cases, new blogger feature vector 'Bfv'
Output: blogger class BC (Yes/No)
Method:
list L1 ←Null //List to store the set of previous solutions of similar problems having 80% similar features with current problem
list L2 ←Null //List to store the set of previous solutions of similar problems having 60% similar features with current problem
Rf ← train_RandomForest(CR) //Random Forest model training based on CR
while test instance t having Bfv is coming
    L1, L2 = Retreive (CR, Bfv)
if (L1!=NULL)
    BC ←Reuse(L1, L2, Rf, Bfv, CR)
else
    BC ←Revise(CR, Bfv, Rf)
end if
end while
```

The training data (from aforesaid blogger dataset) having '*n*' number of cases (blogger feature vectors) is kept in CR in the form of problem-solution pair. Each problem solution pair consists of six attributes where five attributes represents blogger features namely Degree, Political Caprice, Topic, LMT, and LPSS and sixth attribute is the output class of a blogger (Professional or non-professional). FIGURE 3 provides a glimpse of CR.

A. RETRIEVE

Abdelwahed et al. [43] found that the extraction of the best similar cases is tricky because the performance of a classifier

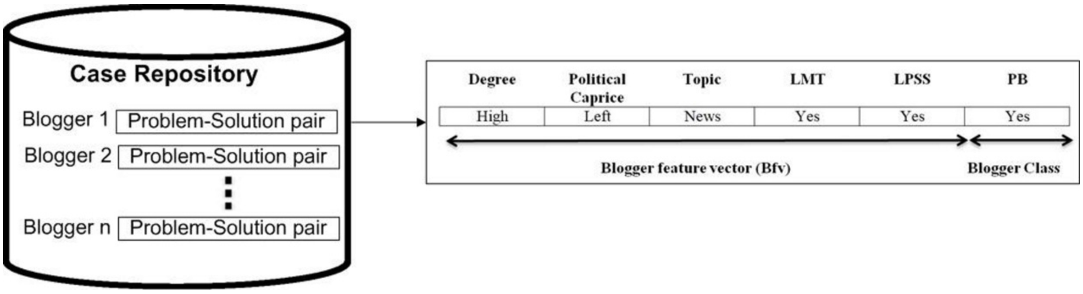

Case Repository
Blogger 1 Problem-Solution pair
Blogger 2 Problem-Solution pair
Blogger n Problem-Solution pair
Degree
Political Caprice
Topic
LMT
LPSS
PB
High
Left
News
Yes
Yes
Yes
Blogger feature vector (Bfv)
Blogger Class

FIGURE 3. Structure of case repository.

is directly affected by the selected similar case solution. This phase is responsible for comparing the current case (new problem) with the existing cases in CR and provides the most matching cases. In the past a number of similarity measures have been used to perform this task, but in this work, we have used aforementioned well-known similarity calculation measures to find the similarity between the new problem and the cases of CR. Algorithm 3 is developed for the retrieve phase of CBR to extract the matching cases of bloggers to predict the new blogger whether he/she is professional or not.

While traversing through all the available cases, if we find 80% or above similarity between new problem and any of existing cases of CR, then list L1 is maintained for storing the output classes of all matching cases of CR. If we find 60% or above but less than 80% similarity between new problem and any of the existing cases, then list L2 is maintained for storing the output classes of all matching cases of CR. The Retrieve phase returns both lists.

Algorithm 3 Pseudocode for Retrieve Phase

```
Input: Given a case repository 'CR' having 'n' cases,
blogger feature vector 'Bfv'
Output: list L1, list L2
Method:
for all training cases tc in CR
    if similarity (tc, Bfv) >= 0.80
        maintain list L1
    else if similarity (tc, Bfv) >= 0.60
        maintain list L2
    end If
end for
return L1,L2
```

B. REUSE

In this phase, solutions of the matching cases are used to devise solution of the current problem. A number of different solution algorithms have been used previously for this phase such as arithmetic average, fuzzy inference rules, and probabilistic models. In this work, we have proposed an algorithm (see Algorithm 4) for the reuse phase. If the length of list L1 is greater than the length of list L2, which means that more similar cases have 80% similarity with the new blogger feature vector (new problem) then, we can reuse the solutions of these cases to devise the solution of the new case. For this purpose, if a most frequently occurring class returned by Algorithm 5 is in odd number in list L1, then it will be taken as a solution to the new problem. However, if there is a case in which 50% solutions of list L1 belong to one class and 50% solutions belong to another class, then one more similar solution is added to list L1 to make odd the possible number of solutions. The first solution of List L2 is used for this purpose and is added to the list L1 provided the list L2 is not empty. Afterwards, the most frequently occurring solution is extracted to predict the solution of the new problem. As an illustration, suppose we have to devise the solution of the following problem as shown in FIGURE 4 and the values in both lists are as follows:

Let L1 = {Yes,Yes,No,No,No} and L2 = {Yes,No} are maintained by retrieve phase while matching the above problem with the available cases in CR. Here, the length of L1 (five elements) is greater than L2 (two elements), so the output class variable (BC) will be assigned the most common value in list L1 which is ''No''. It indicates that if the blogger has the features as shown in FIGURE 4 then he will be non-professional blogger.

On the other hand, if L1 = {Yes,No,Yes,No} then it is clear that both classes ''Yes and No'' are occurring in the same number. Here, we add the first similar solution of L2 i.e. ''Yes'' in the list L1. Now, the list L1 = {Yes,No,Yes,No,Yes}, where the output class variable (BC) will be assigned the most common value in the list L1 i.e. ''Yes''. It points out that if the blogger has the features as shown in FIGURE 4 then he will be a professional blogger. In case, if the length of list L2 is greater than list L1, which shows that the found number of matching cases have 60% or above similarity with the new problem, then Revise (Algorithm 6) will be called.

We have developed Algorithm 5 for finding the frequently occurring solution (majority class) in the list L1 to reuse it.

C. REVISE

In adaptation phase, if list L1 is found empty (in Algorithm 2), it shows that no similar case is found in

Algorithm 4 Pseudocode for Reuse Phase

Input: list ‘L1’, list ‘L2’, RandomForest model ‘Rf’, new blogger feature vector ‘Bfv’, case repository ‘CR’
Output: blogger class BC (Yes/No)
Method:

```
if (len (L1) > len (L2)) //If there are more solutions with 80% matching
    BC← Most_common (L1) // extract most frequent solution
if (len(L1) /2 == count of BC in L1) //if obtained solution is occurring as half elements of L1
        if (L2!=NULL)
            add L2[0] in L1 // use first solution of L2 to make number of solutions odd in L1
            BC← Most_common (L1)
        else
            BC← Most_common (L1)
else
    BC←Revise(CR, Bfv, Rf)
end if
return BC
```

Algorithm 5 Pseudocode for Most_Common

Input:list ‘L1’
Output:most frequently occurring solution Sol
Method:

```
Sol_count←0
Sol←NULL
for each s in L1
    if (L1.count(s) > Sol_count)
    Sol←s
    Sol_count=L1.count(s)
return Sol
```

CR which is 80% similar to new problem or if the list L2 has more elements in it than list L1, which shows that most of the found similar cases have 60% similarity with the new problem, then in both cases, already trained RF model (based on ‘n’ cases of case repository) is used for finding solution and the test case is predicted accordingly. Algorithm 6 is used for this purpose. A number of ML techniques have already been investigated for blogger classification with RF classifier (an ensembling technique), IB1 Classifier (a lazy learning technique) and ANN classifier (an adaptive technique) have been observed to have performed better in classifying bloggers as professional or non-professional. The ANN and IB1 were the candidates to be selected for the revise phase of IB-CBR, but RF beats its competitors in the revise phase. Moreover, Retain phase is called here to store the predicted solution in CR for future use.

Algorithm 6 Pseudocode for Revise Phase

Input: case repository ‘CR’, new blogger feature vector ‘Bfv’, RandomForest model ‘’Rf’
Output: blogger class BC (Yes/No)
Method:

```
BC← Rf.predict (Bfv)
Retain (CR, Bfv, BC)
return BC
```

D. RETAIN

After suggesting the solution to the new problem, this problem-solution pair can be retained in CR. The immediate availability of presently solved problem in CR makes the CBR approach more effective for solving similar cases in future, once their solution is known in solving new problems. Algorithm 7 has been used to store the new problem with its solution in the CR.

Algorithm 7 Pseudocode for Retain Phase

Input: case repository CR, new blogger feature vector Bfv, blogger class BC
Method:

```
CR = CR U{<Bfv, BC>} // store solved problem with its solution is stored in CR
n = n + 1
```

V. EXPERIMENTAL DESIGN AND DISCUSSION

In this section, the details of experimental setup are provided, and the results of the proposed algorithm are discussed. All the experiments were carried out on Windows 10, Intel ®Processor Core^{TM} i7-7500U CPU @ 2.70GHz, 1TB Hard Disk, and 8GB installed RAM. We have implemented the proposed algorithm in python 2.7. The outcomes of IB-CBR are evaluated by using 10-fold cross validation to show its efficiency for the identification of influential bloggers. Moreover, the results of the suggested algorithm are compared with the previously used machine learning techniques such as RF and IB1, ANN, and C4.5 that have performed well for similar problems on the same blogger dataset. The experimental results are evaluated by using standard performance measures. Furthermore, the performance of IB-CBR is examined for different online datasets.

To evaluate the adaptability of the proposed model for blogger classification, the values of performance measures before and after adaptation are recorded. It is found that the accuracy of IB-CBR without adaptation (by using Algorithm 1) is lesser as compared to after adaptation (by using Algorithm 2) for all similarity measures. It indicates that adaptation algorithm has good performance in solving unseen problems. Furthermore, as discussed earlier that the selection of similarity measure is an important decision for the performance of similarity-based classifiers. So, to find the appropriate similarity measure for solving our problem, we found that the Euclidean distance similarity measure

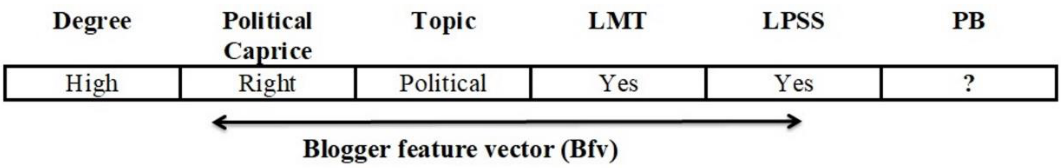


FIGURE 4. An example of a test case.

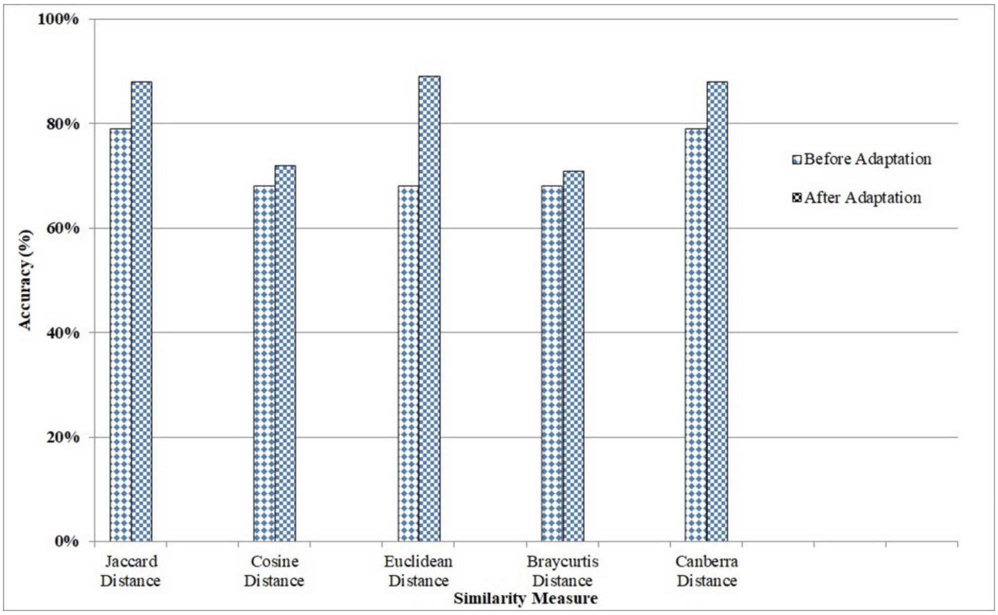


FIGURE 5. Adaptation comparison through IB-CBR in terms of accuracy for influential blogger identification.

showed the best accuracy as 89% after adaptation as shown in FIGURE 5. Also, Jaccard distance similarity and Canberra distance similarity both have achieved 88% accuracy. Whereas Braycurtis distance similarity and Cosine distance similarity have shown 71% and 72% accuracy which is quite lower than the former similarity measures.

Likewise, before adaptation the results for the Cosine, Euclidean, and Braycurtis distance similarity are found similar in terms of all performance measures which do not show the superiority of any particular similarity measure over the other. However, the Jaccard has shown the best value for F-measure as 86% and Canberra has shown the least value for FP Rate as 34% as compared to their competitors.

On the other hand, the results of the IB-CBR model with adaptation for Euclidean distance similarity are observed to be the best in terms of FP Rate, and F-measure as 14%, and 92% respectively as shown in Table 2. However, both of Jaccard, and Canberra distance similarity have shown better TP Rate (94%) as compared to the Euclidean distance similarity. The results of the FP Rate with the most of the other similarity measures are found higher than 14%, which pinpoints that respective measures are less effective than Euclidean distance similarity in identifying actual professional bloggers.

In general FP Rate is decreased and F-measure is increased after adaptation in the case of each similarity measure which shows the strength of the proposed model. However, the similar performance of Jaccard distance similarity and Canberra distance similarity in terms of all performance measures as well as Euclidean distance similarity in terms of accuracy and TP Rate (as shown in FIGURE 5 and Table 2) motivated us to further explore their performance for IB-CBR and to select that similarity measure which can maximally enhance the performance of the proposed classifier.

To further investigate the performance of aforementioned similarity measures, we have obtained iterative adaptive results up to three iterations of IB-CBR. It can be seen in Table 3 given below that the values of all performance measures are not increasing in the case of Euclidean distance similarity. It means that it will not contribute towards the learning capability of IB-CBR in the future classifications.

On the other hand, the Jaccard distance similarity and Canberra distance similarity have shown better results for performance measures in terms of Accuracy, TP Rate, FP Rate, F-measure, and ROC AUC in the second and third iteration. However, as compared to the former similarity measures, Canberra distance similarity maximally enhanced

TABLE 2. Performance comparison for blogger classification by using different similarity measures.

Similarity measure	Before Adaptation (%)			After Adaptation (%)		
	TP Rate	FP Rate	F-measure	TP Rate	FP Rate	F-measure
Jaccard Distance	94	62	**86**	**94**	**36**	**91**
Cosine Distance	100	100	80	100	88	82
Euclidean Distance	100	100	80	92	**14**	**92**
Braycurtis Distance	100	100	80	99	87	82
Canberra Distance	90	**34**	85	**94**	**36**	**91**

TABLE 3. Performance measures comparison for blogger classification by using different similarity measures with one, two and three iterations.

Similarity Measure	No. of iterations	Accuracy	TP Rate	FP Rate	F-measure	ROC AUC
Euclidean Distance	One iteration	89%	92%	14%	92%	90%
	Two iterations	88%	91%	16%	90%	90%
	Three iterations	88%	91%	16%	90%	90%
Jaccard Distance	One iteration	88%	94%	36%	91%	83%
	Two iterations	90%	96%	25%	93%	90%
	Three iterations	92%	95%	14%	94%	97%
Canberra Distance	One iteration	88%	94%	36%	91%	91%
	Two iterations	92%	95%	19%	94%	93%
	Three iterations	**95%**	**97%**	**11%**	**96%**	**98%**

the performance of IB-CBR on third iteration. Canberra distance similarity has achieved an increase in accuracy from 88% to 95%, a gain in TP Rate from 94% to 97%, increasing F-measure from 91% to 96%, and a rise in ROC area under the curve from 91% to 98%, which is higher than other similarity measure candidates. The value of ROC AUC in all iterations is above 90% and approaching to 100%, which shows Canberra contributes towards the excellency of IB-CBR. It shows that the selection of Canberra distance similarity for matching similar cases upon the arrival of a new case will enable IB-CBR to correctly classify bloggers into professional or otherwise based on historical data/experience. Also, FP Rate is found decreasing from 36% to 11% with increasing iterations of IB-CBR, which shows that using this similarity measure, IB-CBR is less likely to classify non-professional bloggers as professional.

The reason behind these findings is that as we are retaining problem-solution pairs in CR after revising in the Revise phase whenever a new problem (with 60% or above similarity with previous cases) is seen, it can effectively help IB-CBR model in the prediction of unseen bloggers into professional or otherwise.The performance of similarity measures is related to the dimensions of data whether it is low or high [44]. High dimensional data can disturb the performance of a classifier due to the fact that more features in the data are less likely to provide new information about the output class; becoming a hurdle in designing a good classifier [45]. Also, a classifier may start looking for patterns to identify the class in irrelevant features rather than the relevant features, hence learning may not be generalizable. This way, it seems that data dimensions are related to the performance of a similarity measure.

Canberra is capable of providing accurate results in the case of high dimensional datasets (having more than two or three dimensions) [44]. In this research work, it is quite possible that better performance of IB-CBR classifier based on Canberra distance similarity is due to high dimensional dataset of bloggers (having six dimensions). Moreover, this similarity measure is likely to achieve more gain in the performance of IB-CBR model than Jaccard distance similarity and Euclidean distance similarity as per increase in the number of iterations.

It is concluded that Jaccard distance is a good similarity measure which can produce an optimal partition having low dimensional spaces in case of high dimensional space [46]. It is a measure of commonness which can calculate the similarity as well as dissimilarity of instances. It measures similarity between two instances by dividing their intersection set with their union set. It can represent similarities between cases belonging to similar class in such a way that can clearly differentiate instances belonging to a different class. It seems easy for Jaccard distance similarity to clearly show the difference between new problem and cases in CR with respect to the class to which a new problem belongs. On the other hand, Euclidean distance measure has been considered preferable in low dimensional data [47]. In our research, Euclidean distance has shown good performance at first, but no performance gain on more iterations of IB-CBR which is probably due to high dimensional input data.

In Algorithm 3, we have used two thresholds 80% and 60% for similarity finding in the case of reusing and revising solution respectively. The selection of these thresholds is based on a series of experiments where we have checked

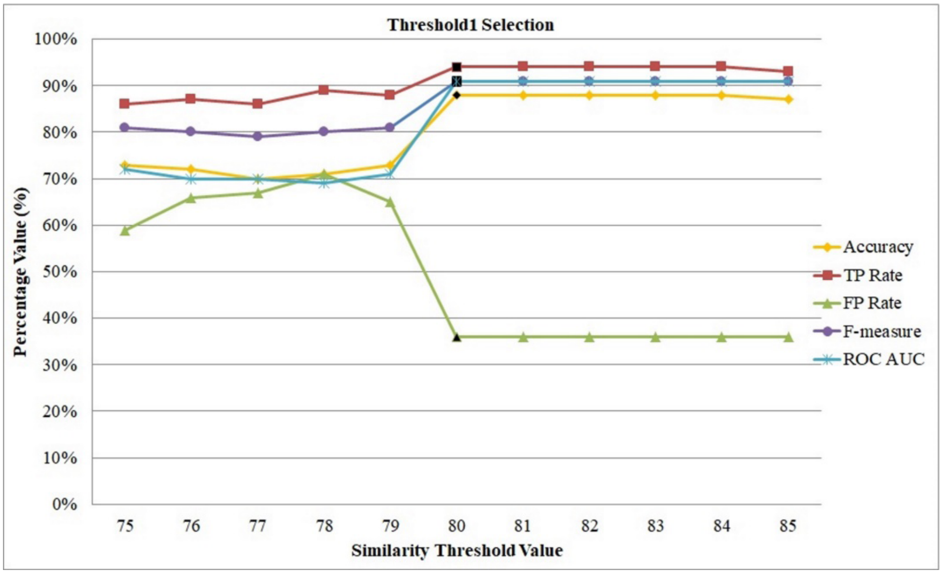


FIGURE 6. Selection of Threshold1 (80 %)based on performance measures for reusing solution where threshold2 i.e. 60% is kept constant.

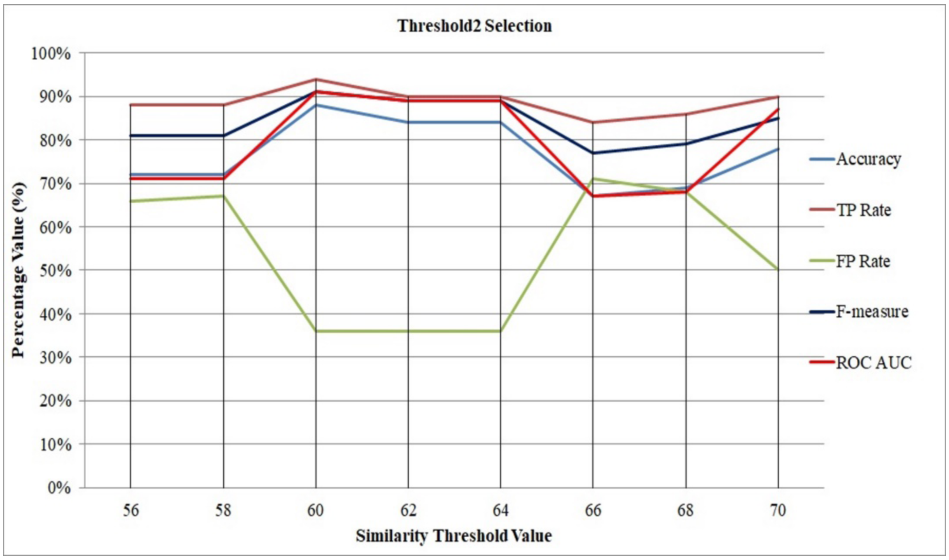


FIGURE 7. Selection of Threshold2 based on performance measures for reusing solution where threshold1 i.e. 80% is kept constant.

the output of IB-CBR by varying both threshold values using the Canberra distance similarity measure. Let's assume that threshold1 = 80% and threshold2 = 60% (as mentioned in Algorithm 3), then by varying threshold1 within the range of 75%-85% similarity and by keeping the threshold2 constant (60%), the results show that on the 80 % threshold value, results start improving for performance measures as shown in the following FIGURE 6 by black marker fill. It shows that if we have a new problem which is 80% similar to any previously seen cases then we can advise solution for that problem by reusing the solution of previous cases.

Similarly, we have seen the performance of IB-CBR for blogger classification by keeping threshold1 constant and by varying threshold2 within the range of 56%-70%. The results shown in FIGURE 7 highlight that upon 60% similarity value (threshold2), the performance of IB-CBR in terms of accuracy, TP Rate, F-measure, and ROC AUC has increased maximally and FP Rate is found minimum. It shows that revising the solution upon finding 60% similarity of a new problem with previously stored cases positively contributes to the performance of IB-CBR.

As it is discussed earlier, we have performed prediction by using RF algorithm for revise phase of IB-CBR model for

TABLE 4. Performance measures comparison of split criteria by using different similarity measures.

Similarity Measures	RF Classifier (Gini Gain)				RF Classifier (Information Gain)			
	Accuracy	TP Rate	FP Rate	F-measure	Accuracy	TP Rate	FP Rate	F-measure
Jaccard Distance	88%	**94%**	36%	**91%**	87%	**94%**	40%	**91%**
Cosine Distance	**72%**	**100%**	**88%**	**82%**	**72%**	**100%**	**88%**	**82%**
Euclidean Distance	89%	92%	14%	92%	84%	91%	40%	89%
Braycurtis Distance	**71%**	**99%**	**87%**	**82%**	**71%**	**99%**	**87%**	**82%**
Canberra Distance	88%	94%	36%	91%	84%	91%	40%	89%

TABLE 5. Performance measures comparison of different classifiers used in the revise phase of IB-CBR.

Similarity measure	RF Classifier (%)				IB1 Classifier (%)				ANN Classifier (%)			
	TP Rate	FP Rate	F-measure	ROC AUC	TP Rate	FP Rate	F-measure	ROC AUC	TP Rate	FP Rate	F-measure	ROC AUC
Jaccard	**94**	**36**	**91**	83	89	66	81	71	87	67	80	70
Cosine	100	88	82	81	100	88	82	76	100	88	82	76
Euclidean	92	14	92	90	94	17	92	78	94	21	92	79
Braycurtis	99	87	82	80	97	93	80	79	97	93	80	79
Canberra	**94**	**36**	**91**	**91**	89	66	81	76	88	73	79	77

adaptation if there is 60% or above similarity between the new problem and previously stored cases in CBR. We have used leading classifiers highlighted by [30] and [33] that can perform well for blogger classification to examine their performance when merged with CBR. We devised a solution to the new problem (if 60% or above but less than 80% matching is found) on the basis of the efficiency of these algorithms for blogger classification. During the revise phase, it can be seen in FIGURE 8 that most of the time, RF classifier outperforms ANN classifier and IB1 classifier by achieving more accuracy in the case of each similarity measure. It shows that RF can be comparatively better merger with CBR.

Likewise, the performance of RF classifier for Jaccard, Braycurtis and Canberra distance similarity in terms of TP Rate, and F-measure is found better as compared to IB1 and ANN classifier in the revise phase of the IB-CBR model. However, FP Rate is found very high in the case of all similarity measures except the Euclidean distance similarity for both IB1 and ANN classifiers. Whereas, in the case of Euclidean distance similarity, although TP Rate has increased in IB1 and ANN than RF but FP Rate is also found increasing. Though, the results of Cosine similarity and Braycurtis similarity in terms of TP Rate are higher than other similarity measures in the case of all classifiers, but, they are also incapable of identifying professional bloggers correctly due to higher FP Rate. The results show that merging of RF with CBR produces greater than 80% ROC area under the curve in case of all similarity measures except Euclidean and Canberra (90% and 91% respectively). It denotes that RF contributes as a good as well as an excellent merger in IB-CBR. On the other hand, IB1 and ANN achieved less than 80% results for ROC AUC, which shows that make IB-CBR as a mediocre classifier.

Besides, we have also investigated the effect of node split criteria such as Gini impurity as well as Entropy with respect to performance metrics. It is found that Gini Gain is found a little better than Information Gain in terms of performance improvement. Table 4 shows that in the case of Gini impurity split criteria, the similarity measures such as Euclidean, and Canberra has shown higher values for Accuracy, TP Rate, and F-measure as compared to entropy. Likewise, FP Rate is found lower in this case, which shows the superiority of Gini impurity over Entropy. On the other hand, when we use Gini impurity as split criteria in the case of Jaccard similarity, only accuracy (such as 88%), and FP Rate (such as 36%) is found a little bit higher and lesser respectively, while TP Rate and F-measure remain same. In both splitting methods, maximally achieved ROC AUC is 91% in the case of Canberra. However, Cosine and Braycutis similarity remain same for both split criteria in terms of performance measures. It is clear from the results that mostly there is no significant differences found in both split criteria while building a decision tree which is consistent with the results discussed by [48]. Due to a minor improvement in results, we have selected Gini impurity as split criteria for RF in further experiments. So, it can be said that overall the results of RF classifier seem better than IB1 classifier and ANN classifier in the revise phase and it should be merged with CBR approach. Table 5 highlights these results.

We have also performed experiments for revise phase with the selected Canberra distance similarity, by using different k-fold cross validation as shown in FIGURE 9. It can be seen that the accuracy, TP Rate, F-measure, and

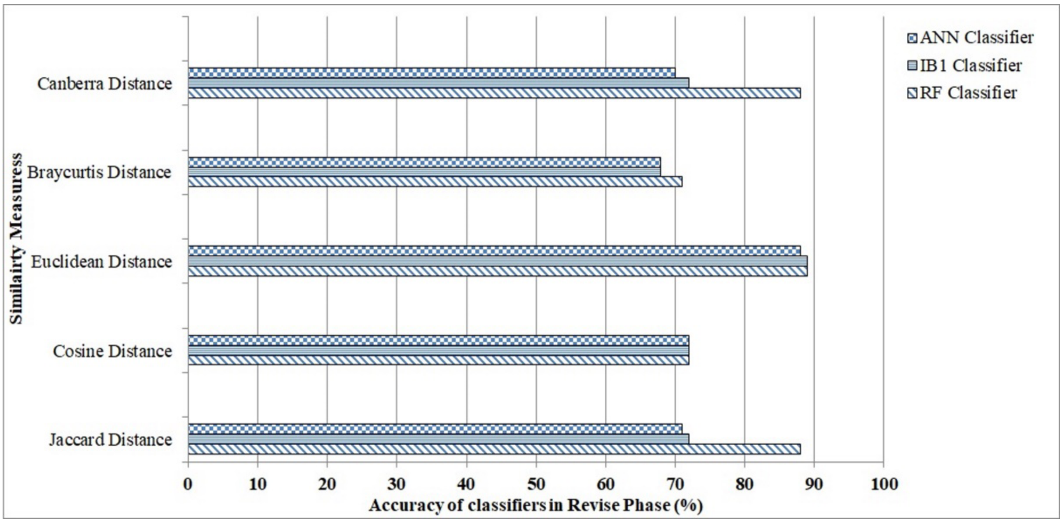


FIGURE 8. Accuracy comparison of RF classifier, IB1 classifier, and ANN classifier used in the revise phase of IB-CBR.

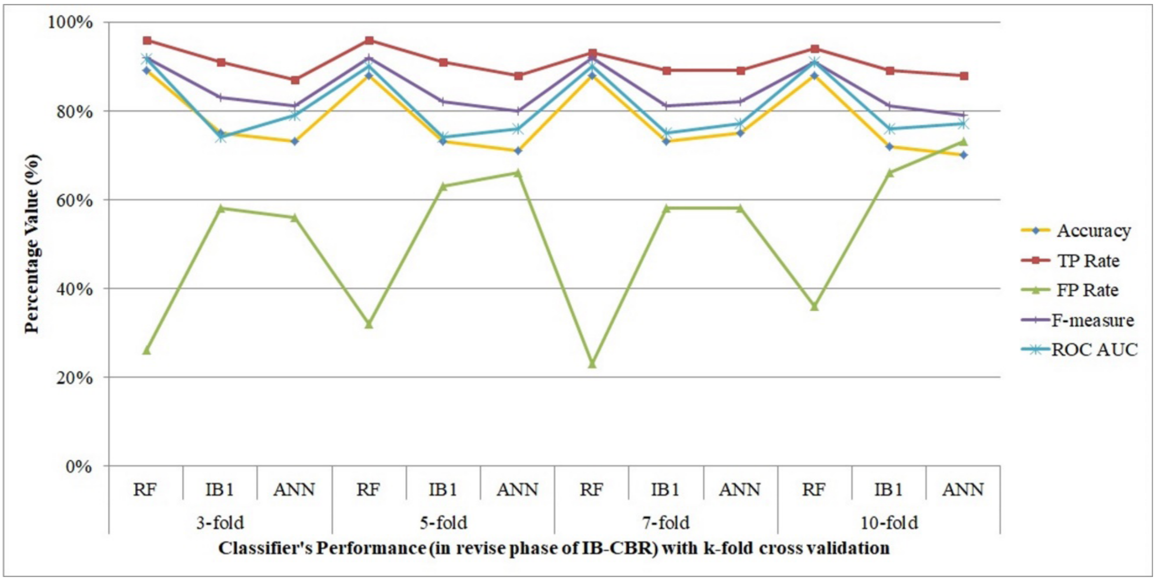


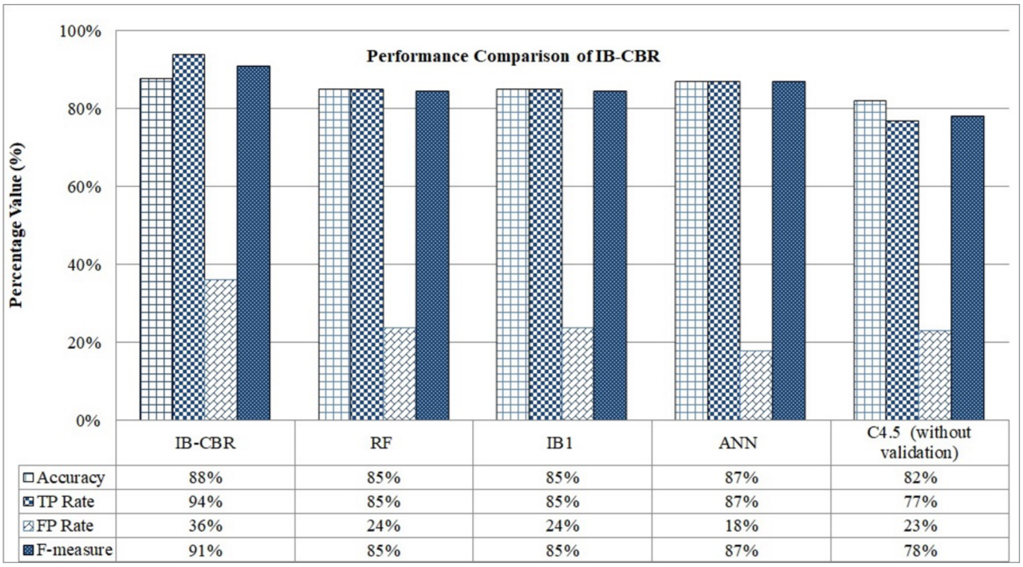
FIGURE 9. Performance comparison of classifiers in revise phase with varying number of folds in cross validation.

ROC AUC obtained by the RF algorithm is higher than the other classifiers used in the revise phase. It shows that if we use RF in Revise phase of the IB-CBR algorithm then it can predict the professional/non-professional bloggers more accurately and can distinctly classify both types of bloggers into two classes. Moreover, FP Rate of RF classifier is found minimum as compared to IB1 classifier and ANN classifier. It shows that RF classifier is less likely to predict non-professional bloggers as a professional which shows that it is highly specific in blogger classification.

In addition, it is found that the IB-CBR algorithm outperforms previous studies by achieving 88% accuracy which have been conducted for professional blogger classification by using the same dataset as shown in FIGURE 10. The TP Rate for the proposed algorithm is 94%, which is relatively high as compared to its competitors. It shows that the proposed algorithm is highly sensitive in the correct prediction (classification) of positive examples (professional bloggers) as positive (professional) and negative examples (non-professional bloggers) as negative (non-professional). Also, the value of F-measure for IB-CBR is found higher, i.e. 91% than the other classifiers which indicates that it can assure that each output class has clear-cut examples such as professional or non-professional bloggers after classification.

A. PROOF OF CONCEPT

Besides blogger dataset, we have investigated the performance of the IB-CBR model on other standard datasets



	IB-CBR	RF	IB1	ANN	C4.5 (without validation)
□ Accuracy	88%	85%	85%	87%	82%
▩ TP Rate	94%	85%	85%	87%	77%
▨ FP Rate	36%	24%	24%	18%	23%
■ F-measure	91%	85%	85%	87%	78%

FIGURE 10. Performance comparison of IB-CBR with RF, IB1, ANN, C4.5 for blogger classification.

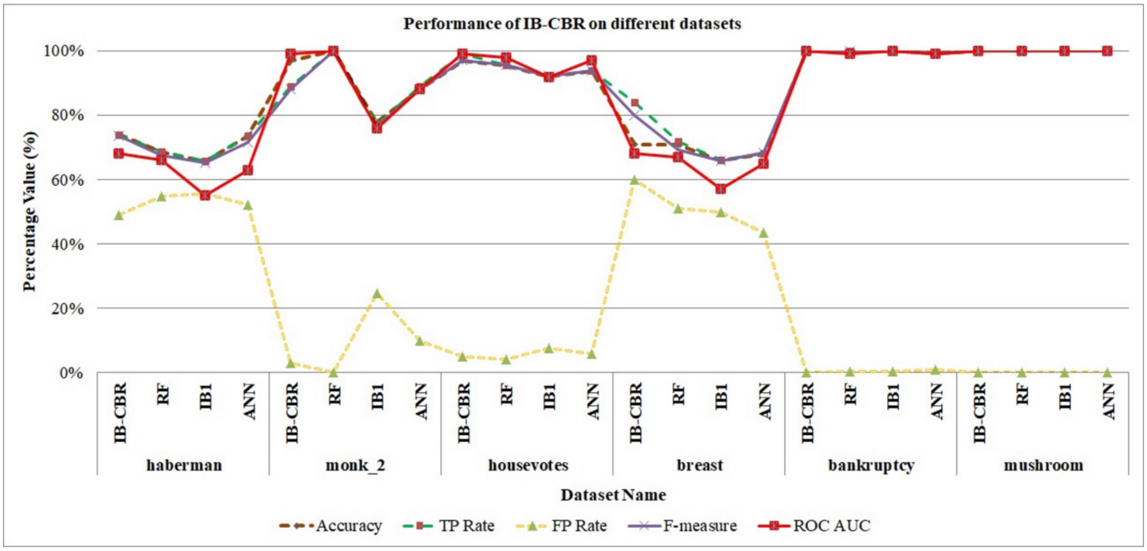


FIGURE 11. Performance of IB-CBR algorithm for prediction.

for its predictive abilities using 10-fold cross validation. These datasets are named as follows: haberman dataset, monk_2 dataset, housevotes dataset, breast dataset, mushroom dataset, that are available online[2] and bankruptcy dataset.[3] We have compared the performance of IB-CBR with RF classifier, IB1 classifier, and ANN classifier (we have compared the results of IB-CBR with previously underlined, and well-performed techniques [30], [33] on these datasets. All the datasets have different number of attributes (high dimensional data) and a varying number of instances. There are no missing values in these datasets and all of them have two output classes.

The results can be seen in FIGURE 11 which show that IB-CBR has achieved 74% accuracy, 74% TP Rate, 49% FP Rate, 73.6% F-measure, and 68% ROC AUC for haberman dataset. On the other hand, it is found that competitive classifiers have lesser accuracy, TP Rate, F-measure, ROC AUC and higher FP Rate than IB-CBR which shows the strength of the proposed model for classification. Similarly, IB-CBR outperforms RF classifier, IB1 classifier, and ANN classifier in terms of performance measures in the case of housevotes dataset, breast dataset, and bankruptcy dataset. In the case of mushroom dataset, all classifiers achieved 100% results in terms of accuracy, TP Rate, F-measure, ROC AUC as

[2] http://sci2s.ugr.es/keel/category.php?cat=clas#sub2

[3] https://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy

well 0% FP Rate, which indicates that this dataset consists of such instances which help in clearly identifying positive instances as positive and vice versa.

However, in the case of monk_2 dataset, RF classifier is at the top w.r.t its performance, following IB-CBR classifier which beats IB1 classifier and ANN classifier. In such a case, both thresholds can be tuned and by increasing the number of iterations of IB-CBR classifier can also help to increase the performance measures. Also, there is a chance that another similarity measure can provide better classification results for this dataset. It is because, the selection of similarity measure that can provide the best separation in prediction analysis may vary from dataset to dataset having differing dimensions [44]. It indicates that providing the best similarity measure in general for all kinds of datasets is not possible.

Being an ensemble method, RF is capable of producing multiple decision trees with respect to same training data instead of making predictions based on a single decision tree. It randomly selects features for the production of the forest of several decision trees which maximizes its ability to beat single classifiers. Its ability to use multiple starting points (such as local optimum solutions) to find an unknown function, strengthens it to opt for best hypothesis. It puts effort to investigate different local optimum solutions because of generating loads of trees [30]. It splits nodes of a decision tree based on two split criteria such as 'Gini impurity' and 'Entropy'. The former specifies the probability of mistakenly categorizing a data point in the dataset. If every datapoint falls in the same class (i.e. only professional or non-professional blogger), then Gini purity will be zero (lowest one) and Gini gain would be maximized. Equation (5) is used to find out Gini impurity.

$$Gini.impurity = \sum_{x=1}^{n} p_x\,(1 - p_x) \tag{5}$$

where n shows the number of classes (in our case there are two classes), and p_x represents the probability for randomly

selecting an element of class x. The Gini gain is determined by the subtraction of weighted impurities of the branches of decision trees from the original impurity. During the process of constructing a decision tree, the best split is picked by minimizing the Gini impurity which ultimately shows maximization of Gini gain.

The latter split criteria is Entropy, which is also a common way to measure the uncertainty in data. Reducing this uncertainty, gives rise to information gain, which highlights the worth of information carrying an attribute for data classification. The attribute with high information gain is selected for splitting a node. Equation (6) is used to calculate the information gain.

$$\text{I.G. (A, B)} = \text{Entropy (A)-Entropy (A,B)} \tag{6}$$

where the first expression represents the entropy of original collection A and the second expression denotes the possible value of entropy after the partitioning of A by selecting attribute B. It can also be seen as an expected reduction in entropy given a particular attribute is selected for splitting the node. Equation (7) can be used to calculate this expected reduction in entropy $I.Gain(A, B)$ by knowing the value of attribute B.

$$I.G.(A, B) = Entropy(A) - \sum_{v \in values(B)} \frac{|A_v|}{|A|} Entropy(A_v) \tag{7}$$

where first expression represents entropy of original collection A and the second expression is the expected value of entropy when A is partitioned by using attribute B. Equation (8) is used to determine entropy.

$$Entropy = \sum_{x} -p_x \log_2 p_x \tag{8}$$

where p_x represents the probability of class x.

A minor variation in the results of IB-CBR in the case of Gini impurity and Entropy is probably due to the working nature of both criteria for splitting a node to construct a decision tree. In the case of former splitting criteria, while producing a decision tree, all the data of the class with the maximum purity is kept in the left sub-tree and all the remaining classes to right sub-tree. On the other hand, the latter breaks the classes into two disjoint subsets and serves to balance the sample size in both sub-trees [49]. However, in each case, the goal is to reduce impurity in data to properly categorize data. This is probably the main reason of good performance of RF.

Though IB1 algorithm can perform the best in the case of related features [38], but as we are using it for prediction when the similarity between the new problem and cases in CR is found from 60% to 80%, which shows that there is less relevancy of attributes. Moreover, instead of building any explicit classification model to classify new problems, it inherits instance-based learning which opts for local approximations. Perhaps, these reasons are behind the lesser performance of IB1 when merged with CBR.

ANN builds a model based on the training data and keeps on reducing the error between the predicted output and targeted output by updating weights and minimizing the error using gradient descent until a reasonable performance of the network is achieved. Equation (9) represents the total error E over the network output units.

$$E(\vec{w}) = 0.5 \sum_{d \in D} \sum_{k \in output\_units} (t_{kd} - p_{kd})^2 \tag{9}$$

where t_{kd} and p_{kd} denotes the targeted and predicted output values with respect to kth output unit and training example d. The testing instances are classified by using this model where weights are already obtained; not updated again. There is a probability that by using these weights may give rise to the error between targeted and predicted outputs which may affect the performance of ANN for future predictions.

In this study, the results of ANN and IB1 are found worse than RF in the revise phase of IB-CBR. The reason for RF performing better could be its nature where not just one but many classifiers work together to decide the fate of an unseen instance. Earlier, a very large comparative study used 121 datasets for classification and 179 classifiers [50] which also supports our observation. It is found that RF performed the best in more than 90% of the cases. On the other hand, ANN produces worse results on average as compared to the former. Moreover, based on the nature of output classes i.e yes/no, we have chosen majority voting for suggesting solution of a new problem which is already used by [26] in the case of nominal data

VI. CONCLUSION AND FUTURE WORK

This study provides an adaptive influential blogger prediction model based on the CBR approach in combination with RF algorithm. This way marketing boards can take benefit by hiring such leading individuals in order to achieve their promotion goals. They can use the influence of influential bloggers and pay them for their services. This study tends to devise a system to search for the most professional bloggers as they will definitely be the most influential. The findings of the IB-CBR algorithm were compared with previously used machine learning techniques such as RF, IB1, ANN, and C4.5. Extensive experiments were performed to evaluate the performance of the proposed algorithm. The results show that the suggested approach has performed well as compared to other machine learning techniques for blogger classification and adaptation. Since the IB-CBR algorithm can solve new problems adaptively instead of explicitly training a model for unobserved instances, which makes it suitable and more adjustable in problem solving in identification of professional/ non-professional bloggers. The IB-CBR model can update itself according to unseen data for improvement of its performance.

In future, we aim to propose a novel framework for the identification of influential blogger, which will be capable of using labeled as well as unlabeled data. Further, deep neural networks could be explored for this problem. We are also interested in collecting a new real dataset of bloggers for this purpose.

ACKNOWLEDGMENT

The authors would like to thank Mr. Abdul Majeed from Metis International, Pakistan, for his useful discussion in technical perspective.

REFERENCES

[1] E. Keller and J. Berry, *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. New York, NY, USA: Free Press, 2003.

[2] Y. Yang and G. Xie, "Efficient identification of node importance in social networks," *Inf. Process. Manage.*, vol. 52, no. 5, pp. 911–922, 2016.

[3] X. Zhao, F. Liu, J. Wang, and T. Li, "Evaluating influential nodes in social networks by local centrality with a coefficient," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 2, p. 35, Jan. 2017.

[4] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Washington, DC, USA, Aug. 2003, pp. 137–146.

[5] Y. Zhang, Z. Wang, and C. Xia, "Identifying key users for targeted marketing by mining online social network," in *Proc. IEEE 24th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Apr. 2010, pp. 644–649.

[6] T. Araujo, P. Neijens, and R. Vliegenthart, "Getting the word out on Twitter: The role of influentials, information brokers and strong ties in building word-of-mouth for brands," *Int. J. Advertising*, vol. 36, no. 3, pp. 496–513, 2017.

[7] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, Hong Kong, Feb. 2011, pp. 65–74.

[8] C. Amos, G. Holmes, and D. Strutton, "Exploring the relationship between celebrity endorser effects and advertising effectiveness: A quantitative synthesis of effect size," *Int. J. Adv.*, vol. 27, no. 2, pp. 209–234, 2008.

[9] S.-A. A. Jin and J. Phua, "Following celebrities' tweets about brands: The impact of Twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities," *J. Advertising*, vol. 43, no. 2, pp. 181–195, Apr. 2014.

[10] N. Booth and J. A. Matic, "Mapping and leveraging influencers in social media to shape corporate brand perceptions," *Corporate Commun., Int. J.*, vol. 16, no. 3, pp. 184–191, 2011.

[11] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the spread of misinformation in social networks," in *Proc. 20th Int. Conf. World Wide Web*, Hyderabad, India, Mar. 2011, pp. 665–674.

[12] M. Han, M. Yan, Z. Cai, Y. Li, X. Cai, and J. Yu, "Influence maximization by probing partial communities in dynamic online social networks," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 4, p. e3054, Apr. 2017.

[13] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Paris, France, Jun. 2009, pp. 199–208.

[14] F. Wang, W. Jiang, X. Li, and G. Wang, "Maximizing positive influence spread in online social networks via fluid dynamics," *Future Gener. Comput. Syst.*, vol. 86, pp. 1491–1502, Sep. 2017.

[15] H. Zhu, X. Yin, J. Ma, and W. Hu, "Identifying the main paths of information diffusion in online social networks," *Phys. A, Stat. Mech. Appl.*, vol. 452, pp. 320–328, Jun. 2016.

[16] M. A. M. A. Kermani, A. Aliahmadi, and R. Hanneman, "Optimizing the choice of influential nodes for diffusion on a social network," *Int. J. Commun. Syst.*, vol. 29, no. 7, pp. 1235–1250, May 2016.

[17] J. M. Larson, "The weakness of weak ties for novel information diffusion," *Appl. Netw. Sci.*, vol. 2, no. 1, p. 14, Dec. 2017.

[18] E. K. Mbaru and M. L. Barnes, "Key players in conservation diffusion: Using social network analysis to identify critical injection points," *Biol. Conservation*, vol. 210, pp. 222–232, Jun. 2017.

[19] K. Zhao and A. Kumar, "Who blogs what: Understanding the publishing behavior of bloggers," *World Wide Web*, vol. 16, nos. 5–6, pp. 621–644, Nov. 2013.

[20] H. U. Khan, A. Daud, U. Ishfaq, T. Amjad, N. Aljohani, and R. A. Abbasi, "Modelling to identify influential bloggers in the blogosphere: A survey," *Comput. Hum. Behav.*, vol. 68, pp. 64–82, Mar. 2017.

[21] H. U. Khan and A. Daud, "Finding the top influential bloggers based on productivity and popularity features," *New Rev. Hypermedia Multimedia* , vol. 23, no. 3, pp. 189–206, Sep. 2016.

[22] F. S. Gharehchopogh, S. R. Khaze, and I. Maleki, "A new approach in bloggers classification with hybrid of k-nearest neighbor and artificial neural network algorithms," *Indian J. Sci. Technol.*, vol. 8, no. 3, pp. 237–246, Feb. 2015.

[23] F. S. Gharehchopogh and S. R. Khaze, "Data mining application for cyber space users tendency in blog writing: A case study," *Int. J. Comput. Appl.*, vol. 47, no. 18, pp. 40–46, Jul. 2013.

[24] N. A. Samsudin, A. Mustapha, and M. H. A. Wahab, "Ensemble classification of cyber space users tendency in blog writing using random forest," in *Proc. 12th Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2016, pp. 169–172.

[25] M. A. Al-Garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, G. Mujtaba, and M. U. S. Khan, "Analysis of online social network connections for identification of influential users: Survey and open research issues," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 16–37, Apr. 2018.

[26] M. J. Khan, M. M. Awais, S. Shamail, and I. Awan, "An empirical study of modeling self-management capabilities in autonomic systems using case-based reasoning," *Simul. Model. Pract. Theory*, vol. 19, no. 10, pp. 2256–2275, Nov. 2011.

[27] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, Jan. 2003.

[28] C. Fullwood, K. Melrose, N. Morris, and S. Floyd, "Sex, blogs, and baring your soul: Factors influencing UK blogging strategies," *J. Assoc. Inf. Sci. Technol.*, vol. 64, no. 2, pp. 345–355, Feb. 2013.

[29] B. Quadir and N.-S. Chen, "The effects of reading and writing habits on blog adoption," *Behav. Inf. Technol.*, vol. 34, no. 9, pp. 893–901, Feb. 2015.

[30] Y. Asim, A. R. Shahid, A. K. Malik, and B. Raza, "Significance of machine learning algorithms in professional blogger's classification," *Comput. Electr. Eng.*, vol. 65, pp. 461–473, Jan. 2018.

[31] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[32] J. B. Awotunde and R. G. Jimoh, "A model for identifying influential bloggers using social proof, mining comment, and emerging topics in social media," *Comput. Inf. Syst.*, vol. 23, no. 1, pp. 19–25, Feb. 2019.

[33] Y. Asim, B. Raza, A. K. Malik, S. Rathore, and A. Bilal, "Improving the performance of professional blogger's classification," presented at the Int. Conf. Comput., Math. Eng. Technol., Sukkur, Pakistan, Mar. 2018.

[34] P. SanMiguel and T. Sádaba, "Nice to be a fashion blogger, hard to be influential: An analysis based on personal characteristics, knowledge criteria, and social factors," *J. Global Fashion Marketing*, vol. 9, no. 1, pp. 40–58, 2018.

[35] F. Li and T. C. Du, "Maximizing micro-blog influence in online promotion," *Expert Syst. Appl.*, vol. 70, pp. 52–66, Mar. 2017.

[36] I. Watson, "Case-based reasoning is a methodology not a technology," *Knowl.-Based Syst.*, vol. 12, no. 5, pp. 303–308, Oct. 1999.

[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[38] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[39] T. M. Mitchell, "Artificial neural network," in *Machine Learning*. New York, NY, USA: McGraw-Hill, 1999, pp. 81–127.

[40] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[41] A.-B. M. Salem and T. Shmelova, "Intelligent expert decision support systems: Methodologies, applications," in *Socio-Technical Decision Support in Air Navigation Systems: Emerging Research and Opportunities*. New York, NY, USA: IGI Global, 2018, pp. 215–241.

[42] S. M. F. D. S. Mustapha, "Case-based reasoning for identifying knowledge leader within online community," *Expert Syst. Appl.*, vol. 97, pp. 244–252, May 2018.

[43] M. F. Abdelwahed, M. Saleh, and A. E. Mohamed, "Speeding up single-query sampling-based algorithms using case-based reasoning," *Expert Syst. Appl.*, vol. 114, pp. 524–531, Dec. 2018.

[44] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144059.

[45] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[46] M. Baisantry and D. P. Shukla, "Comparison of different similarity measures for selection of optimal, information-centric bands of hyperspectral images," presented at the Observing Changing Earth, Sci. Decis. Monit., Assessment, Projection, Sioux Falls, South Dakota, 2017.

[47] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. Int. Conf. Database Theory*, London, U.K., Oct. 2001, pp. 420–434.

[48] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, 2004.

[49] L. Breiman, "Technical note: Some properties of splitting criteria," *Mach. Learn.*, vol. 24, no. 1, pp. 41–47, Jul. 1996.

[50] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

YOUSRA ASIM has received the BS degree in software engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan, in 2006 and MSCS degree from Kinnaird College, Lahore, Pakistan in 2008. She has been teaching as Lecture since 2009 and now she is working as Assistant Professor of Computer Science in Govt. College for Women Sihal since 2016. She is a Ph.D. student in the CS department at COMSATS, Islamabad. Her research interests are Social Networks, Influential Nodes, Data Mining, Machine Learning and Privacy.

BASIT RAZA is working as Assistant Professor in the department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. He received his Ph.D. (Computer Science) degree in 2014 from International Islamic University, Islamabad, Pakistan. He has published a number of conference and journal papers of internal repute. His research interests are Database management system, Security and Privacy, Data Mining, Data Warehousing, Machine Learning and Artificial Intelligence.

AHMAD KAMRAN MALIK received his Ph.D. from the Vienna University of Technology (TU-Wien), Austria. He is working as an Assistant Professor at COMSATS University Islamabad (CUI), Islamabad, Pakistan. He studied MS in Computer Science at Muhammad Ali Jinnah University, Islamabad. Since 1999 he has been teaching and supervising computer science students at undergraduate and graduate level. Currently, his research interest is focused on Social Network Analysis, Access Control, and Data Science.

AHMAD R. SHAHAID is currently working as Assistant Professor at COMSATS Institute of Information Technology, Islamabad, Pakistan. He did his PhD in Computer Science from York, UK in 2012. During his PhD he worked on automatically building a WordNet for four languages, namely, English, German, French and Greek. After his PhD, he has been working in the areas of Computer Vision and Pattern Recognition, Machine Learning, and Natural Language Processing. A few of the problems that he has worked on include cancer detection, pedestrian detection, driver fatigue detection, and data mining.

HANI ALQUHAYZ is Assistant Professor at Majmaah University, Saudi Arabia. He is Assistant Professor in Computer Science department in College of Sciences at Majmaah University. Hani has PhD in Computer Science from De Montfort University in UK. Also, got Masters in Information Systems Management and Bachelor of Computer Science. His research interests in Wireless security, scheduling, and image processing.