IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.
www.ijiset.com
ISSN 2348 – 7968

Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News

Sushilkumar Kalmegh

Associate Professor, Department of Computer Science, Sant Gadge Baba Amravati University
Amravati, Maharashtra- 444602, India.

Abstract

The amount of data in the world and in our lives seems ever-increasing and there's no end to it. The Weka workbench is an organized collection of state-of-the-art machine learning algorithms and data pre-processing tools. The basic way of interacting with these methods is by invoking them from the command line. However, convenient interactive graphical user interfaces are provided for data exploration, for setting up large-scale experiments on distributed computing platforms, and for designing configurations for streamed data processing. These interfaces constitute an advanced environment for experimental data mining. Classification is an important data mining technique with broad applications. It classifies data of various kinds. This paper has been carried out to make a performance evaluation of REPTree, Simple Cart and RandomTree classification algorithm. The paper sets out to make comparative evaluation of classifiers REPTree, Simple Cart and RandomTree in the context of dataset of Indian news to maximize true positive rate and minimize false positive rate. For processing Weka API were used. The results in the paper on dataset of Indian news also show that the efficiency and accuracy of RandomTree is good than REPTree, and Simple Cart.

Keywords— *Simple Cart, RandomTree, REPTree, Weka, WWW*

1. Introduction

The amount of data in the world and in our lives seems ever-increasing and there's no end to it. We are overwhelmed with data. Today Computers make it too easy to save things. Inexpensive disks and online storage make it too easy to postpone decisions about what to do with all this stuff, we simply get more memory and keep it all. The World Wide Web (WWW) overwhelms us with information; meanwhile, every choice we make is recorded. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. Lying hidden in all this data is information.

In *data mining*, the data is stored electronically and the search is automated or at least augmented by computer. Even this is not particularly new. Economists, statisticians, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction. What is new is the staggering increase in opportunities for finding patterns in data.

Data mining is a topic that involves learning in a practical, non theoretical sense. We are interested in techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it. Experience shows that in many applications of machine learning to data mining, the explicit knowledge structures that are acquired, the structural descriptions, are at least as important as the ability to perform well on new examples. People frequently use data mining to gain knowledge, not just predictions.

2. Literature Survey

2.1 WEKA

Weka was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka provides implementations of learning algorithms that can be easily apply to dataset. It also includes a variety of tools for transforming datasets, such as the algorithms.

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools. It is designed so that we can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a variety of learning algorithms, it includes a wide range of pre processing tools. This diverse and comprehensive toolkit is

ISSN 2348 – 7968

accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand. All algorithms take their input in the form of a single relational table in the ARFF format. The easiest way to use Weka is through a graphical user interface called Explorer as shown in **figure 1.** This gives access to all of its facilities using menu selection and form filling.

Fig. 1 : Weka GUI Explorer

The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. Advantages of Weka include:

- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets. The Explorer interface features several panels providing access to the main components of the workbench. **Figure 2** shows Opening of file *.arff by Weka Explorer and **Figure 2** shows processing of arff file

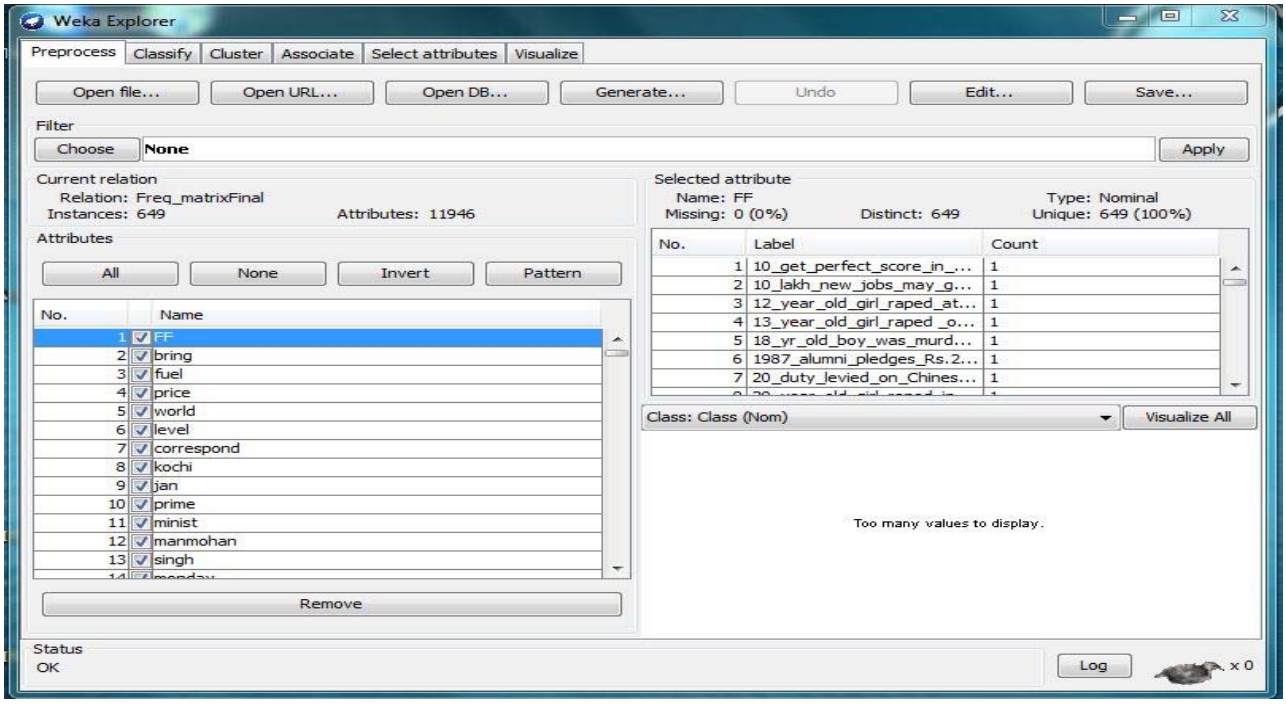

Fig. 2 : Opening Of File *.arff By Weka Explorer

ISSN 2348 – 7968

new data instance. The tree it creates is exactly that: a tree

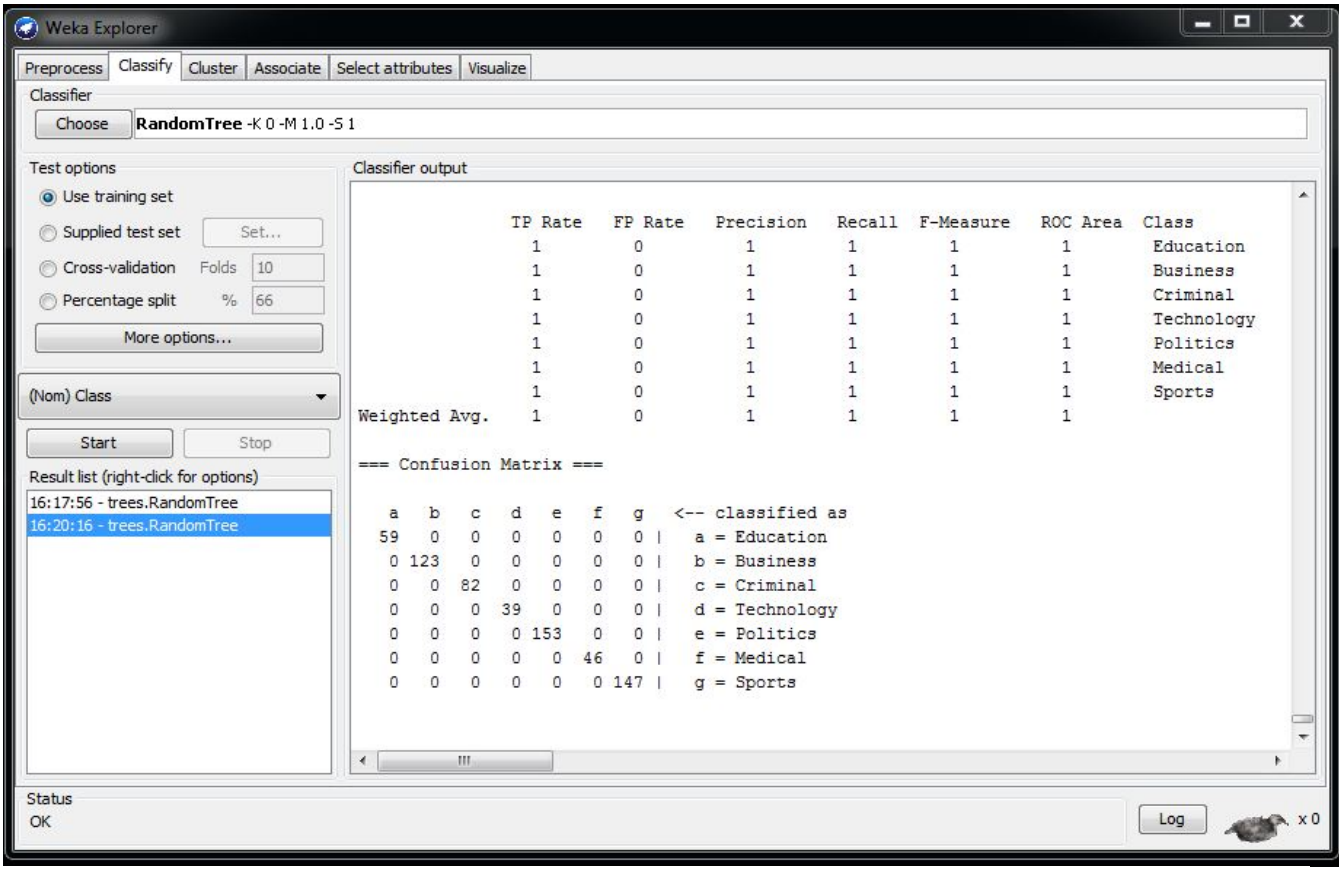

Fig. 3: Processing Of arff File By RandomTree Classifier

2.2 Classification

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity.

Classification is a data mining algorithm that creates a step-by-step guide for how to determine the output of a decision must be made based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output. Sounds confusing, but it's really quite straightforward.

There is also some argument over whether classification methods that do not involve a statistical model can be considered “statistical". Other fields may use different terminology: e.g. in community ecology, the term "classification" normally refers to cluster analysis, i.e. a type of unsupervised learning, rather than the supervised learning. [2].

2.2.1 REPTree

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

ISSN 2348 – 7968

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated for class gender having six possible values. [1] [3] [4]

2.2.2 Simple Cart

Simple Cart method is CART (Classification And Regression Tree) analysis. CART is abbreviated as Classification and Regression Tree algorithm. It was developed by Leo Breiman in the early 1980s. It is used for data exploration and prediction also. Classification and regression trees are classification methods which in order to construct decision trees uses historical data. CART uses learning sample which is a set of historical data with pre-assigned classes for all observations for building decision tree.

Simple Cart (Classification and regression tree) is a classification technique that generates the binary decision tree. Since output is binary tree, it generates only two children. Entropy is used to choose the best splitting attribute. Simple Cart handles the missing data by ignoring that record. This algorithm is best for the training data. Classification and regression trees (CART) decision tree is a learning technique, which gives the results as either classification or regression trees, depending on categorical or numeric data set.

It methodology proposed by is perhaps best known and most widely used. It uses cross-validation or a large independent test sample of data to select the best tree from the sequence of trees considered in the pruning process. The basic CART building algorithm is a greedy algorithm in that it chooses the locally best discriminatory feature at each stage in the process. This is suboptimal but a full search for a fully optimized set of question would be computationally very expensive. The CART approach is an alternative to the traditional methods for prediction. In the implementation of CART, the dataset is split into the two subgroups that are the most different with respect to the outcome. This procedure is continued on each subgroup until some minimum subgroup size is reached. [1] [5] [6] [7]

2.2.3 RandomTree Classifiers

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node.

Random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. Random trees is a collection (ensemble) of tree predictors that is called forest. The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the classifier response is the average of the responses over all the trees in the forest.

Random Trees are essentially the combination of two existing algorithms in Machine Learning: single model trees are combined with Random Forest ideas. Model trees are decision trees where every single leaf holds a linear model which is optimised for the local subspace described by this leaf. Random Forests have shown to improve the performance of single decision trees considerably: tree diversity is generated by two ways of randomization. First the training data is sampled with replacement for each single tree like in Bagging. Secondly, when growing a tree, instead of always computing the best possible split for each node only a random subset of all attributes is considered at every node, and the best split for that subset is computed. Such trees have been for classification Random model trees for the first time combine model trees and random forests. Random trees employ this produce for split selection and thus induce reasonably balanced trees where one global setting for the ridge value works across all leaves, thus simplifying the optimization procedure. [1] [8] [9] [10]

3. System Design

In order to co-relate News with the categories, a model based on the machine learning and XML search was designed. Flow diagram of the model for news resources is shown below in **fig 4.** As a input to the model, various news resources are considered which are available online like the news in Google news repository or online paper like Times of India, Hindustan Times etc. Around 649 news were collected on above repository. In order to extract context from the news and co-relate it with the proper e-content, the News was process with stemming and tokenization on the news contents. The news then was converted into the term frequency matrix for further

ISSN 2348 – 7968

analysis purpose. Based on this data, features (i.e. metadata) were extracted so that contextual assignment of the news to the appropriate content can be done. This process is known as metadata processing in the above flow diagram. Title of the also contains useful information in the abstract form, the title also can be considered as Metadata. The title of the news is processed using NLP

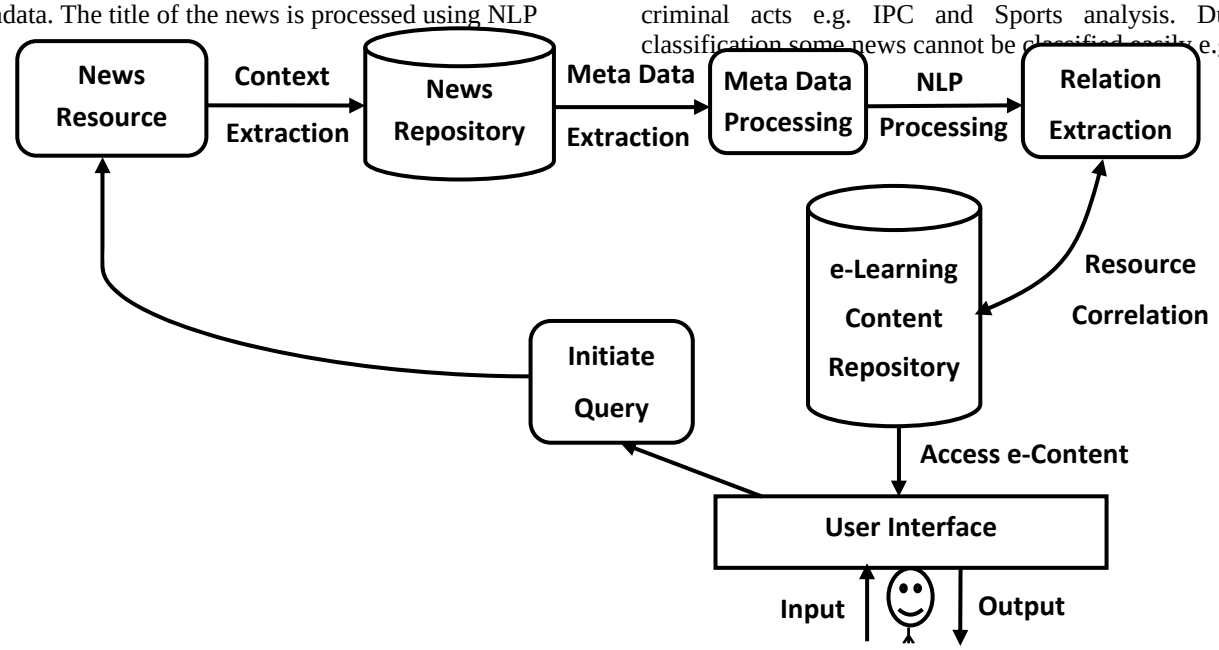

Fig. 4. Flow Diagram Of The Model

libraries (Standford NLP Library) to extract various constituents of it. The output of NLP process was also used to co-relate the News (textual, audio, video) to the concern e-learning contents. This process can be initiated automatically when the user access any content from e-Learning data repository.

As shown in the figure, a news resource is processed to correlate with the e-Contents available. On the similar way, other text resources can be added directly with the e-Content in the repository, Image or Video resource can be processed for meta-data available. And thus can be searched with the related e-Contents. [11]

4. Data Collection

Hence it was proposed to generate indigenous data. Consequently the national resources were used for the research purpose. Data for the purpose of research has been collected from the various news which are available in various national and regional newspapers available on internet. They are downloaded and after reading the news they are manually classified into 7 (seven) categories. There were 649 news in total. The details are as shown in **Table I**.

The attributes consider for this classification is the topic to which news are related; the statements made by different persons; the invention in Business, Education, Medical, Technology; the various trends in Business; various criminal acts e.g. IPC and Sports analysis. During classification some news cannot be classified easily e.g.

(1) Political leader arrested under some IPC code,

(2) Some invention made in medicine and launched in the market & business done per annum.

Hence, there will be drastic enhancement in e-Contents when we refer to the latest material available in this regards. For example, if some e-Content refers to the political situation of India, then the references needs to be dynamic as the situation may change depending on the result of election. [15]

Table I. Categorization Of News

News Category	Actual No. Of News
Business	123
Criminal	82
Education	59
Medical	46
Politics	153

ISSN 2348 – 7968

Sports	147
Technology	39
Total	**649**

5. Performance Analysis

The News so collected needed a processing. Hence as given in the design phase, all the news were processed for stop word removal, stemming, tokenization and ultimately generated the frequency matrix. Stemming is used as many times when news is printed, for a same there can be many variants depending on the tense used or whether it is singular or plural. Such words when processed for stemming, generates a unique word. Stop words needs to be removed as they do not contribute much in the decision making process. Frequency matrix thus generated can be processed for generating a model and the model so generated was used in further decision process. With the model discussed above, three classifier REPTree, Simple Cart and RandomTree were used on the data set of 649 news. For processing Weka APIs were used. The result after processing is given in the form of confusion matrix which is shown in **Table 2, 4 and Table 6.**

REPTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. The performance for Indian News repository has given 0.236% TP and 0.236% FP and area under ROC curve is 0.5%. Simple Cart is a classification technique that generates the binary decision tree. Since output is binary tree, it generates only two children. Entropy is used to choose the best splitting attribute. Simple Cart handles the missing data by ignoring that record. Classification and regression trees (CART) decision tree is a learning technique, which gives the results as either classification or regression trees, depending on categorical or numeric data set. The performance for Indian News repository has given 0.236% TP and 0.236% FP and area under ROC curve is 0.5% same as REPTree. Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. This makes Random Tree to achieve high accuracy. Using Random Tree the performance for Indian News repository has given 100% TP and 0% FP and area under ROC curve is 100%.

It can be seen from following **Table 3, 5 and Table 7**.

Classified as ➡	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	0	0	0	0	**59**	0	0
Business	0	0	0	0	**123**	0	0
Criminal	0	0	0	0	**82**	0	0
Technology	0	0	0	0	**39**	0	0
Politics	0	0	0	0	**153**	0	0
Medical	0	0	0	0	**46**	0	0
Sports	0	0	0	0	**147**	0	0

Table 2: Confusion Matrix for REPTree

Table 3: Table showing True Positive and False Positive Rate of REPTree

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	0	0	0	0	0	0.5
Business	0	0	0	0	0	0.5
Criminal	0	0	0	0	0	0.5

ISSN 2348 – 7968

Technology	0	0	0	0	0	0.5
Politics	1	1	0.236	1	0.382	0.5
Medical	0	0	0	0	0	0.5
Sports	0	0	0	0	0	0.5
Weighted Avg. ➡	0.236	0.236	0.056	0.236	0.09	0.5

Table 4: Confusion Matrix for Simple Cart

Classified as ➡	**Education**	**Business**	**Criminal**	**Technology**	**Politics**	**Medical**	**Sports**
Education	0	0	0	0	**59**	0	0
Business	0	0	0	0	**123**	0	0
Criminal	0	0	0	0	**82**	0	0
Technology	0	0	0	0	**39**	0	0
Politics	0	0	0	0	**153**	0	0
Medical	0	0	0	0	**46**	0	0
Sports	0	0	0	0	**147**	0	0

Table 5: Table showing True Positive and False Positive Rate of SimpleCart

Class ⬇	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Education	0	0	0	0	0	0.5
Business	0	0	0	0	0	0.5
Criminal	0	0	0	0	0	0.5
Technology	0	0	0	0	0	0.5
Politics	1	1	0.236	1	0.382	0.5
Medical	0	0	0	0	0	0.5
Sports	0	0	0	0	0	0.5
Weighted Avg. ➡	0.236	0.236	0.056	0.236	0.09	0.5

Table 6. Confusion Matrix for RandomTree

Classified as ➡	**Education**	**Business**	**Criminal**	**Technology**	**Politics**	**Medical**	**Sports**
Education	**59**	0	0	0	0	0	0
Business	0	**123**	0	0	0	0	0
Criminal	0	0	**82**	0	0	0	0
Technology	0	0	0	**39**	0	0	0

ISSN 2348 – 7968

Politics	0	0	0	0	**153**	0	0
Medical	0	0	0	0	0	**46**	0
Sports	0	0	0	0	0	0	**147**

Table 7. Table showing True Positive and False Positive Rate of RandomTree

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	100%	0%	1	1	1	100%
Business	100%	0%	1	1	1	100%
Criminal	100%	0%	1	1	1	100%
Technology	100%	0%	1	1	1	100%
Politics	100%	0%	1	1	1	100%
Medical	100%	0%	1	1	1	100%
Sports	100%	0%	1	1	1	100%
Weighted Avg.➡	100%	0%	1	1	1	100%

A consolidated performance of the three algorithms used to process News data set can be seen below in **Table 8.**

Table 8. Showing Correct and Wrong Prediction of Classifier.

Classifier ⟶		**REPTree**		**Simple Cart**		**RandomTree**	
News Category	**Actual No. Of News**	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**
Education	**59**	**00**	**59**	**00**	59	**59**	00
Business	**123**	**00**	**123**	**00**	123	**123**	00
Criminal	**82**	**00**	**82**	**00**	82	**82**	00
Technology	**39**	**00**	**39**	**00**	39	**39**	00
Politics	**153**	**153**	**00**	**153**	00	**153**	00
Medical	**46**	**00**	**46**	**00**	46	**46**	00
Sports	**147**	**00**	**147**	**00**	147	**147**	00

ISSN 2348 – 7968

Total	**649**	**153**	496	**153**	496	**649**	00
Percentage ⟶		**23.57**	76.43	**23.57**	76.43	**100**	**00**

6. Conclusions

As per the previous discussion identification of news from dynamic resources can be done with the propose model, we used three classifier i.e. REPTree, Simple Cart and RandomTree to analyze the data sets. As a result it is found that RandomTree algorithm performs best in categorizing all the News. Overall Performance of REPTree and Simple Cart algorithm is not acceptable, because it is seen that both the algorithm are able to classify only politics News correctly. This also can be seen from Table 3, Table 5, Table 7 and Table 8 above.

References

[1] Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier.
[2] http://en.wikipedia.org/wiki/Classification
[3] Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp-155-160
[4] Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8(Version 5), August 2014, pp-91-93
[5] Sunita B. Aher, Lobo L.M.R.J., "COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS", International Journal of Information Technology and Knowledge Management, July-December 2012, Volume 5, No. 2, pp. 239-243
[6] Deepali Kharche, K. Rajeswari, Deepa Abin, "COMPARISON OF DIFFERENT DATASETS USING VARIOUS CLASSIFICATION TECHNIQUES WITH WEKA", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pp- 389-393
[7] S. S. Aman, Kumar Sharma, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," International Journal on Computer Science and Engineering, vol. 3, no. 5, (2011). pp- 1890-1895.
[8] en.wikipedia.org/wiki/Random_tree
[9] Bernhard Pfahringer, "Random model trees: an e_ective and scalable regression method" University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/~bernhard
[10] K. Wisaeng , "A Comparison of Different Classification Techniques for Bank Direct Marketing", International Journal of Soft Computing and Engineering (IJSCE), Volume-3, Issue-4, September 2013, pp-116-119
[11] Sushilkumar. R. Kalmegh, "Successful Assessment of Categorization of Indian News Using JRip and Nnge Algorithm", International Journal of Emerging Technology and Advanced engineering, Volume 4, Issue 12, December 2014 pp- 395-402

Sushilkumar R. Kalmegh received M.Sc. Computer Science (1994), Ph.D. (2014) from Sant Gadge Baba Amravati University, Amravati. Working as Associate Professor in Computer Science in the faculty of Engg. & Tech. in the Department of Computer Science at SGBAU. Amravati. Published 8 Research paper in international Journal. Member of Indian Science, CSI and ISTE. Area of interest is Data Mining.