Effective classification of Indian News using Lazy Classifier IB1And IBk from weka

Sushilkumar R. Kalmegh

Department of Computer Science, Sant Gadge Baba Amravati University, Amravati

sushil.kalmegh@gmail.com

Abstract--- **The amount of data in the world and in our lives increasing day by day with rapid speed. It seems ever-increasing and there's no end to it. We are overwhelmed with data. The WWW overwhelms us with information. 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. Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. It classifies data of various kinds. There are many classification problem occurs in different application areas and need to be solved. Different types of classification algorithms like lazy, tree-based, rule-based, etc are widely used. This paper has been carried out to make a performance evaluation of Lazy Classifier algorithms IB1 and IBk by using different Test Mode. The paper sets out to make comparative evaluation of classifiers IB1 & IBk in test mode (i) evaluate on training data, (ii) 5-fold cross-validation and (iii)10-fold cross-validation in the context of dataset of Indian news to maximize true positive rate and minimize false positive rate.**

Keywords **--- Classification, IB1, IBk, Lazy Classifier, WEKA**

I. INTRODUCTION

Each of the past three centuries has been dominated by a single technology. The eighteenth century was the time of the great mechanical systems accompanying the Industrial Revolution. The nineteenth century was the age of the stream engine. During the twentieth century, the key technology has been information gathering, processing and distribution. Among other developments, we have seen the birth and unprecedented growth of the computer industry. Now as we have entered in the twenty-first century all the most of all manual services are replaced by machine operation i.e. complete computerization. INTERNET has become a major channel of the resources and information. The World Wide Web (WWW) overwhelms us with information; meanwhile, every choice we make is recorded. 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. 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. Text mining, a deviation of data mining is the study of large databases and retrieving interesting patterns or non-trivial information from them. The only difference between text mining and data mining is that in data mining, the tools handle structured data while in text mining, the tools handle unstructured data or semi-structured data from databases. Data mining techniques include clustering, classification, prediction, association rules and sequential patterns.

Here the technique used is Classification. It is used to process a large amount of data and classifies them on the basis of class labels and training set. 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. Classification models entail assigning the data to a predefined category. The main aim of classification is to reduce the classification error. This is a two-step process. In the first step the model is created using classification algorithm on training set. Further, in the second step, the created model is then tested by a predefined test set to measure the accuracy and performance of the model.

In order to get the details of this methodology, this paper is organized into six parts. First part discusses introduction followed by the literature required for analysis of methods implemented. Third one is System Design followed by datasets used for analysis. Fifth is the Performance Analysis and then conclusions.

II. LITERATURE SURVEY

A. 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 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 modelling, 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 pre-processing and modelling techniques.

• Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the a , where each data point is described by a fixed number of a ttribute 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 3** shows processing of arff file for BI1 Classifier(Test Mode : Evaluate on Training Data). [1], [11]

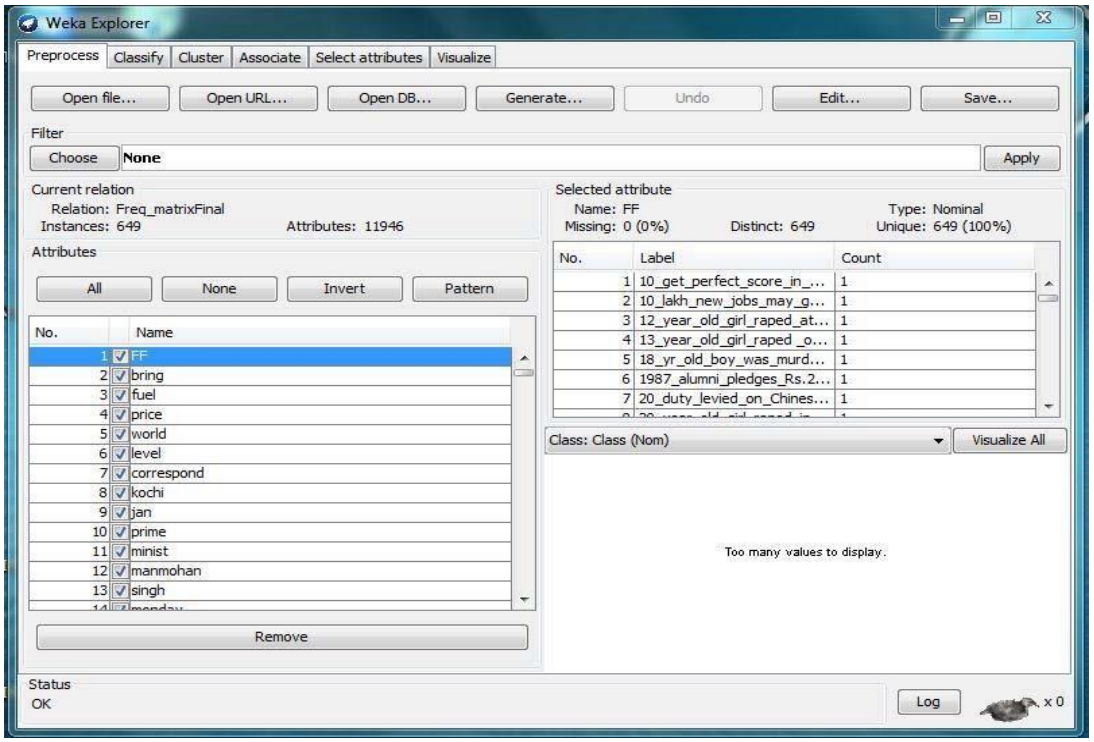

Fig. 2 Opening Of File *.arff By Weka Explorer

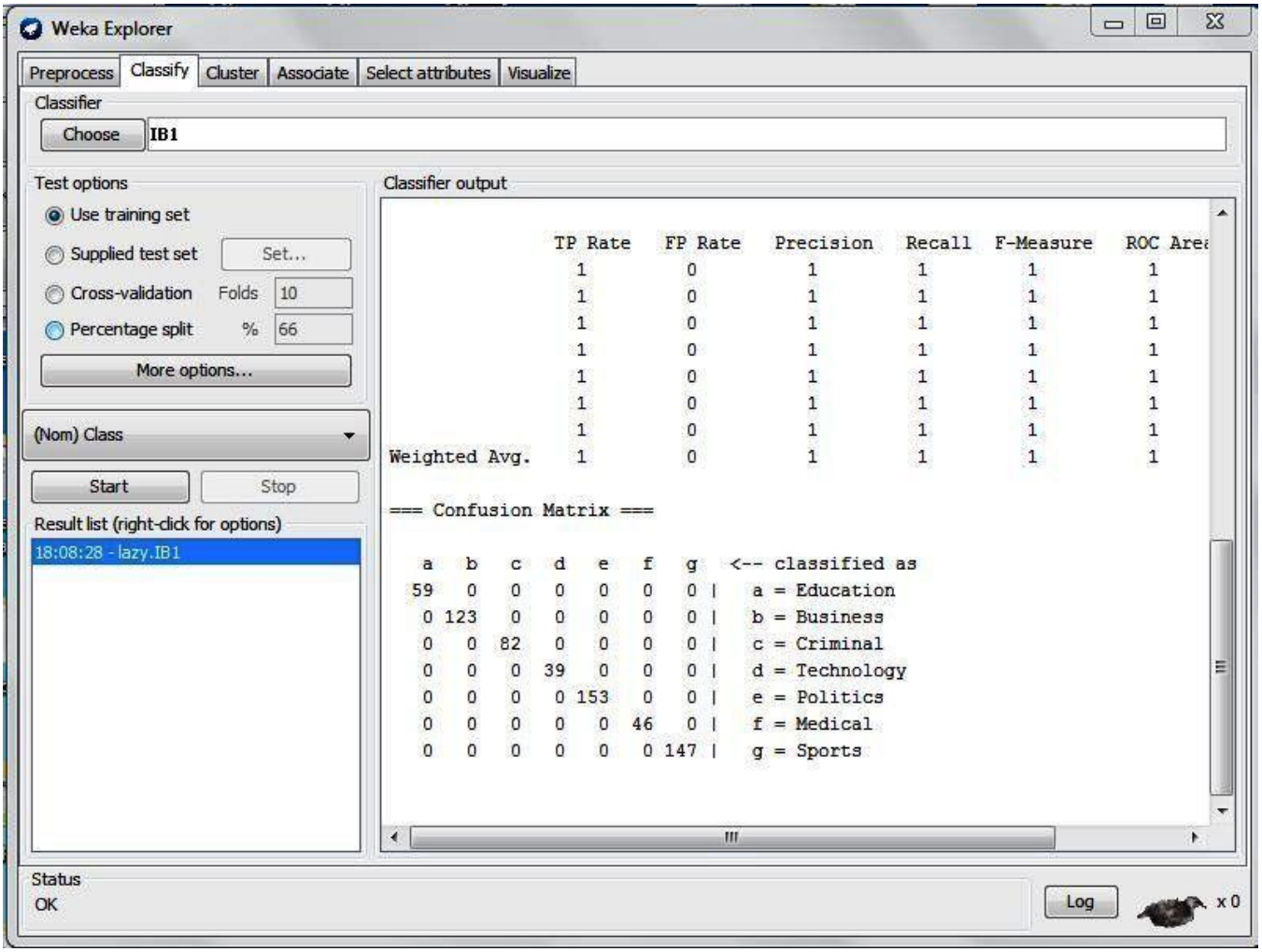

Fig. 3 Processing Of arff File By BI1 Classifier, Test Mode: Evaluate on Training Data

B. 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 new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where 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. [1], [11]

1) *Lazy Classifiers:* Lazy learners store the training instances and do no real work until classification time. Lazy is a classification technique, it includes IB1, IBk, KStar, LWL methods to classify the database. Lazy method doesn’t do anything until last minute. The lazy learners use the same dataset as both training set and testing set.

The main benefit gained in employing a lazy learning method is that the target function will be approximated locally such as in the k-nearest neighbour algorithm. The disadvantages with lazy learning include (1) the large space requirement to store the complete training dataset. (2) lazy learning methods are usually slower to evaluate. Lazy learning solve multiple problems consecutively and deal the problem area in successful. In this paper comparative assessment has been done using IB1 and IBk Lazy Classifiers in test mode (i) evaluate on training data, (ii) 5-fold cross-validation and (iii)10-fold cross-validation in the context of dataset of Indian news. [1], [3], [7]

IB1: IB1 is a basic instance-based learner that finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training instance. If multiple instances have the same (smallest) distance to the test

instance, the first one found is used. IB1 is identical to the nearest neighbour algorithm except that it normalizes its attributes' ranges, processes instances incrementally, and has a simple policy for tolerating missing values.

Nearest neighbour is one of the simplest learning/classification algorithms, and has been successfully applied to a broad range of problems. The nearest neighbour classifier works based on the intuition that the classification of an instance is likely to be most similar to the classification of other instances that are nearby to it. [1], [2], [3], [4], [5], [6]

IBk : IBk is an implementation of the k-nearest-neighbours classifier. In weka it's called IBk (instance based learning with parameter k) and it's in the lazy class folder. Fundamentally "IB" remains for Instance-Based and "k" determines number of neighbors that are analysed. It can select appropriate value of K based on cross-validation. IBk is an instance-based learning approach like the K-nearest neighbour method. The basic principle of this algorithm is that each unseen instance is always compared with existing ones using a distance metric, most commonly Euclidean distance and the closest existing instance is used to assign the class for the test sample weka's default setting is K = 1. Compared to other algorithms, it needs more time to predict the test samples' classes.

Opening of IBk classifier have following steps. The first step to choose weka Explorer initially, then choose dataset, and choose classify tap to get options from IBk implementation. It has the cross-validation option that can help by choosing the best value automatically. Weka uses cross-validation to select the best value for KNN (that is k-nearest neighbor algorithm).

It can also do distance weighting. A variety of different search algorithms can be used to speed up the task of finding the nearest neighbors. The default is the same as for IB1—that is, the Euclidean distance. The number of nearest neighbors (default k = 1) can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation, subject to an upper limit given by the specified value. Predictions from more than one neighbor can be weighted according to their distance from the test instance. [1], [3], [4], [5], [7], [8], [9], [10]

III. System Design

In order to co-relate News with the categories, a model has been designed. Flow diagram of the model for news resources is shown below in **Figure 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, the News was process with Stop words removal, stemming and tokenization on the news contents. The news then was converted into the term frequency matrix for further analysis purpose. The frequency matrix is having extension .csv, so it has to be converted in arff format for processing by WEKA. Based on this data, features (i.e. metadata) were extracted so that contextual assignment of the news to the appropriate content can be done. Title of the news 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 libraries (Standford NLP Library) to extract various constituents of it

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

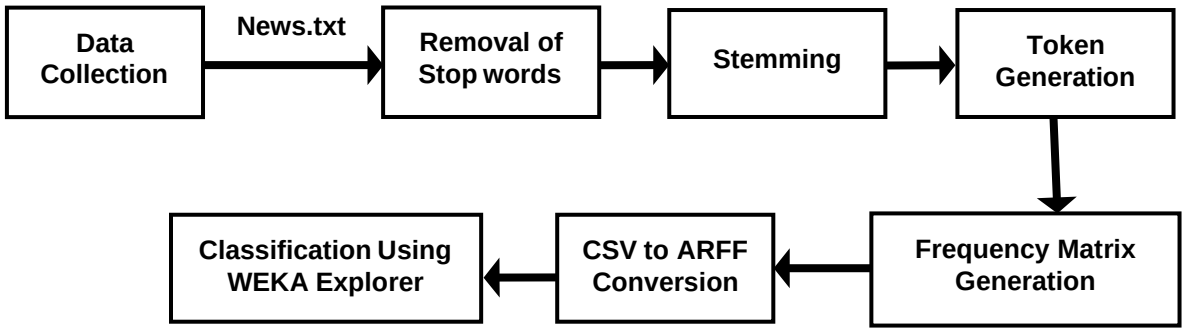

Fig. 4: Flow diagram of the model

IV. 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 1.

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.

TABLE 1
CATEGORIZATION OF NEWS

News Category	Actual No. Of News
Business	123
Criminal	82
Education	59
Medical	46
Politics	153
Sports	147
Technology	39
Total	649

Hence, there will be drastic enhancement in the Contents when we refer to the latest material available in this regards. For example, if some news 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. [11]

V. 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, classifier **IB1** and **IBk** from **Lazy** were used on the data set of 649 news. For processing Weka GUI interface were used. The result after processing data is given in following **Table 2** showing **correctly/incorrectly classified instances** from total 649 instances (i.e. total no. of 649 news). Further the results in the form of confusion matrix for test mode i) evaluate on training data ii) 5-fold cross-validation and iii) 10-fold cross-validation, which are shown in following **Table 3, 5, & 7** by using **classifier IB1** and **Table 9,11 & 13** by using **classifier IBk** respectively. True Positive and False Positive Rate matrix for test mode i) evaluate on training data, ii) 5-fold cross-validation and ii) 10-fold cross-validation which are shown in following **Table 4, 6 & 8** by using **classifier IB1** and **Table 10,12 &14** by using **classifier IBk** respectively.

Overall Performance of **IB1** and **IBk** algorithm is excellent giving **100%** accuracy for test mode i) evaluate on training data, it can been seen from following **Table 3 & 4** by using **classifier IB1** and **Table 9 & 10** by using **classifier IBk**. This may be due to IBk Instance-Based learner with fixed neighborhood. K sets the number of neighbors to use. IB1 is equivalent to IBk for K = 1. WEKA's nearest neighbor implementations (IBk) has been used to generate a classifier based on one neighbor (IB1). IB1 is identical to the nearest neighbour algorithm except that it normalizes its attributes' ranges, processes instances incrementally. The default is the same as for IB1—that is, the Euclidean distance. The number of nearest neighbors (default k = 1) can be specified explicitly.

However from the following **Table 5, 6, 7 & 8** by using **classifier IB1** and **Table 11, 12, 13 & 14** by using **classifier IBk** for test mode ii) 5-fold cross-validation and ii) 10-fold cross-validation most of the news from all category are classified into other category. Minute observation of these tables shows that maximum news from category **Business** and **Politics** are correctly classified. This is because every category has some or other references of the other category. Hence as it can be seen in the **Table 3 & 4** and **Table 9 &10** it has given 100% accuracy for Test mode: evaluate on training data. But this 100% accuracy is not achieved for Test mode: 5-fold cross-validation and 10-fold cross-validation. The another reason for this is that, in n-fold cross-validation, the original sample is randomly partitioned into n subsamples. Of the n subsamples, a single subsample is retained as the validation data for testing the model, and the remaining n – 1 subsamples are used as training data. The cross-validation process is then repeated n times (the folds), with each of the n subsamples used exactly once as the validation data. The n results from the folds then can be averaged (or otherwise combined) to produce a single estimation. [1], [11]

Table 2
Table Showing Correctly/Incorrectly Classified Instances From Total 649 Instances (Total No. Of News)

Lazy Classifier ➡	IB1			IBk		
Test Mode ➡	Evaluate on Training Data	5-fold cross-validation	10-fold cross-validation	Evaluate On Training Data	5-fold cross-validation	10-fold cross-validation
Correctly Classified Instances	649 (100%)	221 (34.0524%)	214 (32.9738%)	649 (100%)	221 (34.0524%)	214 (32.9738%)
Incorrectly Classified Instances	0 (0%)	428 (65.9476%)	435 (67.0262%)	0 (0%)	428 (65.9476%)	435 (67.0262%)

Table 3
Confusion Matrix For Lazy.IB1 For Test Mode : Evaluate On Training Data

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
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 4
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IB1 FOR TEST MODE : EVALUATE ON TRAINING DATA

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

TABLE 5
CONFUSION MATRIX FOR LAZY.IB1 FOR TEST MODE : 5-FOLD CROSS-VALIDATION

Classified as ➡	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	4	36	7	0	10	0	2
Business	0	100	13	0	7	0	3
Criminal	0	40	25	0	17	0	0
Technology	0	28	6	0	4	0	1
Politics	0	72	12	0	69	0	0
Medical	0	36	5	0	3	0	2
Sports	0	88	5	0	31	0	23

TABLE 6
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IB1 FOR TEST MODE : 5-FOLD CROSS-VALIDATION

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	6.8%	0%	100%	6.8%	12.7%	53.4%
Business	81.3%	57%	25%	81.3%	38.2%	62.1%
Criminal	30.5%	8.5%	34.2%	30.5%	32.3%	61%
Technology	0%	0%	0%	0%	0%	50%
Politics	45.1%	14.5%	48.9%	45.1%	46.9%	65.3%
Medical	0%	0%	0%	0%	0%	50%
Sports	15.6%	1.6%	74.2%	15.6%	25.8%	57%
Weighted Avg.➡	34.1%	15.7%	46.5%	34.1%	29.4%	59.2%

TABLE 7
CONFUSION MATRIX FOR LAZY.IB1 FOR TEST MODE : 10-FOLD CROSS-VALIDATION

Classified as ➡	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	4	45	3	0	5	0	2
Business	0	118	0	0	4	0	1
Criminal	0	55	17	0	10	0	0
Technology	0	35	0	0	4	0	0
Politics	0	95	4	0	54	0	0
Medical	0	42	0	0	3	0	1
Sports	0	108	1	0	17	0	21

TABLE 8
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IB1 FOR TEST MODE : 10-FOLD CROSS-VALIDATION

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	6.8%	0%	100%	6.8%	12.7%	53.4%
Business	95.9%	72.2%	23.7%	95.9%	38%	61.8%
Criminal	20.7%	1.4%	0.68%	20.7%	31.8%	59.7%
Technology	0%	0%	0%	0%	0%	50%
Politics	35.3%	8.7%	55.7%	35.3%	43.2%	63.3%
Medical	0%	0%	0%	0%	0%	50%
Sports	14.3%	0.8%	84%	14.3%	24.4%	56.7%
Weighted Avg.➡	33%	16.1%	54.3%	33%	28.1%	58.4%

TABLE 9
CONFUSION MATRIX FOR LAZY.IBK FOR TEST MODE : EVALUATE ON TRAINING DATA

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
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 10
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IBK FOR TEST MODE : EVALUATE ON TRAINING DATA

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

TABLE 11
CONFUSION MATRIX FOR LAZY.IBK FOR TEST MODE : 5-FOLD CROSS-VALIDATION

Classified as ➡	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	4	36	7	0	10	0	2
Business	0	100	13	0	7	0	3
Criminal	0	40	25	0	17	0	0
Technology	0	28	6	0	4	0	1
Politics	0	72	12	0	69	0	0
Medical	0	36	5	0	3	0	2
Sports	0	88	5	0	31	0	23

TABLE 12
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IBK FOR TEST MODE : 5-FOLD CROSS-VALIDATION

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	6.8%	0%	100%	6.8%	12.7%	53.3%
Business	81.3%	57%	25%	81.3%	38.2%	61.7%
Criminal	30.5%	8.5%	34.2%	30.5%	32.3%	63%
Technology	0%	0%	0%	0%	0%	51%
Politics	45.1%	14.5%	48.9%	45.1%	46.9%	66.1%
Medical	0%	0%	0%	0%	0%	50.2%
Sports	15.6%	1.6%	74.2%	15.6%	25.8%	58.9%
Weighted Avg.➡	34.1%	15.7%	46.5%	34.1%	29.4%	60%

TABLE 13
CONFUSION MATRIX FOR LAZY.IBK FOR TEST MODE : 10-FOLD CROSS-VALIDATION

Classified as ➡	Education	Business	Criminal	Technology	Politics	Medical	Sports
Education	4	45	3	0	5	0	2
Business	0	118	0	0	4	0	1
Criminal	0	55	17	0	10	0	0
Technology	0	35	0	0	4	0	0
Politics	0	95	4	0	54	0	0
Medical	0	42	0	0	3	0	1
Sports	0	108	1	0	17	0	21

TABLE 14
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF LAZY.IBK FOR TEST MODE : 10-FOLD CROSS-VALIDATION

Class ⬇	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Education	6.8%	0%	100%	6.8%	12.7%	52.9%
Business	95.9%	72.2%	23.7%	95.9%	38%	61.6%
Criminal	20.7%	1.4%	68%	20.7%	31.8%	59.7%
Technology	0%	0%	0%	0%	0%	51.2%
Politics	35.3%	8.7%	55.7%	35.3%	43.2%	62.3%
Medical	0%	0%	0%	0%	0%	50.6%
Sports	14.3%	0.8%	84%	14.3%	24.4%	57.2%
Weighted Avg.➡	33%	16.1%	54.3%	33%	28.1%	58.3%

VI. CONCLUSIONS

This paper has designed a model which will help to categorize the lazy classifier **IB1** and **IBk** from WEKA in different test mode (i) evaluate on training data (ii) 5-fold cross-validation and (iii)10-fold cross-validation in the context of dataset of Indian news

As per the previous discussion identification of news from dynamic resources can be done with the propose model. As a result it is found that **IB1** and **IBk** algorithm performs well in categorizing all the News for Test mode: evaluate on training data. This is due to IBk Instance-Based learner with fixed neighborhood. K sets the number of neighbors to use. IB1 is equivalent to IBk for K = 1. WEKA's nearest neighbor implementations (IBk) has been used to generate a classifier based on one neighbor (IB1). Overall Performance of **IB1** and **IBk** algorithm is not acceptable for the test mode: 5-fold cross-validation and 10-fold cross-validation, except maximum news from category **Business** and **Politics** are correctly classified. For overall data set detection rate (True Positive rate) for **IB1** and **IBk** clasiifier is 100% for the test mode : evaluate on training data and whereas it is 34.1% for the Test mode: 5-fold cross-validation and surprisingly 33% for the Test mode: 10-fold cross-validation.

ACKNOWLEDGMENT

Author thanks to Dr. Sachin N. Deshmukh, Professor, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad who initiated this line of research; authorities of Sant Gadge Baba Amravati University, Amravati for giving an opportunity to publish the paper. Author also thanks University of Waikato for WEKA tool availability as an open source. Finally author thanks to the entire researcher, whose papers were used as a reference as listed in references section.

REFERENCES

[1] Ian H. Witten, Eibe Frank and Mark A. Hall, *Data Mining Practical Machine Learning Tools and Techniques, Third Edition*, Morgan Kaufmann Publishers is an imprint of Elsevier, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA, 2011.

[2] Jasmina Novakovic and Sinisa Rankov, ―Classification Performance Using Principal Component Analysis and Different Value of the Ratio R‖, *Int. J. of Computers, Communications & Control*, Vol. VI, No. 2, pp. 317-327, June 2011

[3] Ms S. Vijayarani and Ms M. Muthulakshmi, ―Comparative Analysis of Bayes and Lazy Classification Algorithms‖, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, Issue 8, pp. 3118-3124, August 2013

[4] R. Preethi, G. M. SuriyaaKumar, N. G. Bhuvaneswari Amma and G Annapoorani, ―Performance Analysis of Classifiers to Effieciently Predict Genetic Disorders Using Gene Data‖, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, Issue 11, pp. 6960-6966, November 2014

[5] V. Prasathkumar, A. Deepalakshmi and N.Ramkumar, ―Performance Investigation of Lazy Classifiers for the Classification of Multivariate Data‖, *Journal of Information technology and Its Applications*, Volume 2, Issue 2, pp. 1-9, 2017.

[6] Jasmina NOVAKOVIĆ, Perica STRBAC and Dusan BULATOVIĆ, ―TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS‖, *Yugoslav Journal of Operations Research*, 21, Number 1, pp. 119-135, 2011

[7] K.K.Revathi and K.K.Kavitha, ―COMPARISON OF CLASSIFICATION TECHNIQUES ON HEART DISEASE DATA SET‖, *International Journal of Advanced Research in Computer Science*, Volume 8, No. 9, pp. 276-280, November-December 2017

[8] Govinda.K and Narendra B,, ―Opinion mining using Classification Techniques‖, International Journal of Pure and Applied Mathematics, Volume 118, No. 9, pp. 535-544, 2018

[9] S. Venkata Lakshmi and T. Edwin Prabakaran, ―Performance Analysis of Multiple Classifiers on KDD Cup Dataset using WEKA Tool‖, *Indian Journal of Science and Technology*, Vol 8(17), pp. 1-10, August 2015

[10] Sonali Kadam, Rutuja Pawar, Manisha Kumari, Shweta Phule and Priyansha Kher, ―Performance Analysis of Pre-Processing Techniques with Ensemble of 5 Classifiers‖, *International Journal on Recent and Innovation Trends in Computing and Communication*, Volume: 5, Issue: 5, pp. 1250 – 1255, May 2017

[11] Sushilkumar Rameshpant Kalmegh, ―Analysis of Classification Method ―VFI‖ from WEKA by Using Different Test Mode‖, *IJISET - International Journal of Innovative Science, Engineering & Technology*, Vol. 3 Issue 4, pp. 258-265, April 2016.