Definite Detection of Categorization of Indian News by Using PART and FT Algorithm

Sushilkumar R. Kalmegh
Associate Professor, Department of Computer Science, Sant Gadge Baba Amravati University
Amravati (M.S.), India.
sushil_kalmegh@rediffmail.com
Sachin N. Deshmukh
Associate Professor, Department of Comp. Engg. & I.T., Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (M.S.), India.
sndeshmukh@hotmail.com

ABSTRACT

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Recent developments of e-learning specifications such as Learning Object Metadata (LOM), Sharable Content Object Reference Model (SCORM), Learning Design and other pedagogy research in semantic e-learning have shown a trend of applying innovative computational techniques, especially Semantic Web technologies, to promote existing content-focused learning services to semantic-aware and personalised learning services. This paper has been carried out to make a performance evaluation of PART and FT classification algorithm. The paper sets out to make comparative evaluation of classifiers PART and FT 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 FT is good than PART.

KEYWORDS

FT, LOM, LTSA, PART, SCORM, Weka API

1 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 and hence released human intelligence is utilized in further developments.

INTERNET has become a major channel of the resources and information. All most all e-activities are based on the Internet. Among these e-Activities, e-Learning is the fastest growing activity as it has huge market potential. e-Learning encompasses the collection of different resources like Text, Image, Video, Audio, Web-Links etc. This inclusion makes it different than the web pages. As the latest stage of learning and training evolution, e-Learning is supposed to provide intelligent functionalities not only in processing multi-media education resources but also in supporting context-sensitive pedagogical education processes.

Searching the Internet using the keywords for the expected text may be a simple task but integration of other multimedia resources which are semantically relevant to the e-Learning topic becomes a difficult task unless and until specific keywords are assigned manually. Being e-Learning as an organized information and knowledge base, it needs proper inclusion of available multimedia resources. This inclusion is generally based on the keywords. However, if a semi automatic methodology designed, the task of proper inclusion of the resource can be done efficiently.

Considering the drawbacks and accordingly the need of further development in the area of e-Learning, this paper proposes a model for contextual linking of dynamic information from web to be linked to the contents designed for e-Learning. This dynamicity will held the learner get references of current activities online without human intervention. The process of dynamic inclusion is described here with the example of News data which published daily and hence dynamic.

In this case, there is a need of a design of a framework which can integrate dynamic multimedia content to the existing e-contents. This paper discusses the methodology for such integration. In order to get the details of this methodology, this paper is organized into five parts. First part discusses the concept of e-learning followed by the literature required for analysis of methods implemented. Third one discusses the technique of classification. Fourth one is System Design followed by datasets used for analysis. Sixth is the Performance Analysis and then conclusions.

2 e-LEARNING

e-learning is a new education concept by using the Internet technology, it deliveries the digital content, provides a learner-orient environment for the teachers and students. The e-learning promotes the construction of life-long learning opinions and learning society.

It means: e-learning is a new education concept; it may differ from the old educational concept. Delivery of the digital content is the main characters of e-learning. This definition extends the environment on the Internet. It means that the Internet provides a learning environment for the students and teachers. This environment is learner-oriented, so we can throw out the thoughts of traditionally teacher-center's instruction in classroom. As a new concept of education, e-learning gives a condition for us to realize the life-long learning principle and help us to build a more real learning society. e-learning plays a major role in high education for the reason of fast need of high education.

The e-learning industry refers to the effective integration of a range of technologies across all areas of learning, e-learning technologies are designed to support learning by encompassing a range of media, tools, and environments. It allows for both synchronous and asynchronous learning environments. e-learning acts as a catalyst for authentic and meaningful learning experiences [1].

3 LITERATURE SURVEY

The major implementation that includes the intelligence in e-Learning is ConKMEL. To resolve the knowledge integration and management problem in multimedia e-Learning, it has proposed a semantic context aware approach, which features an integrated contextual knowledge management framework to support intelligent e-Learning.[2]

Traditional web-based e-learning systems use a web browser as the interface. Through run-time learning environments (either compatible or incompatible with SCORM) [3] users could access the learning objects, which are directly linked to multimedia learning resources such as lecture video/audio, presentation slides and reference documents. However, the limitation in generic specification support does not affect their compliancy with e-Learning standards such as SCORM (Shareable Content Object Reference Model) [3] for content management and LOM (Learning Object Model) [4] for learning object description. A flow in traditional e-Learning system is given in **Fig 1**.

Weihong Huang et. al. [5] has proposed an intelligent semantic e-Learning framework which presents semantic information processing, learning process support and personalized learning support issues in an integrated environment. Architecture of the above framework is as given below in **Fig 2**. In addition to the traditional learning information flow, three new components namely semantic context model, intelligent personal agents and conceptual learning theories are introduced to bring in more intelligence. Intelligent personal agents perform adequate personal trait information profiling and deliver personalized learning services. Semantic context model uses semantic information for static resource and dynamic process retrieves information from WWW and the future Semantic Web, referring to ontologies or knowledge bases.

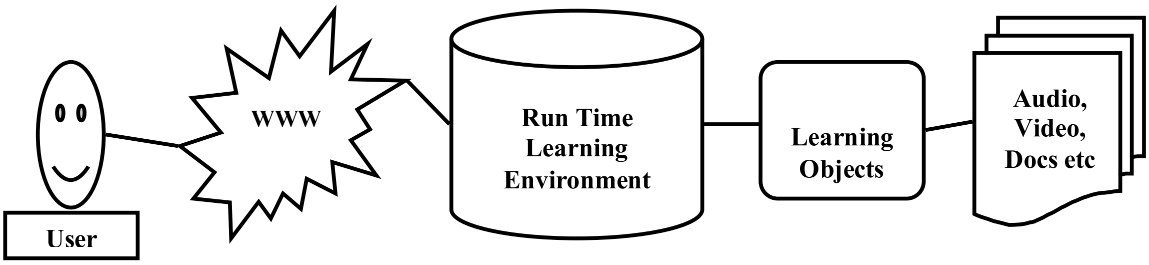

Figure 1. Traditional e-Learning System

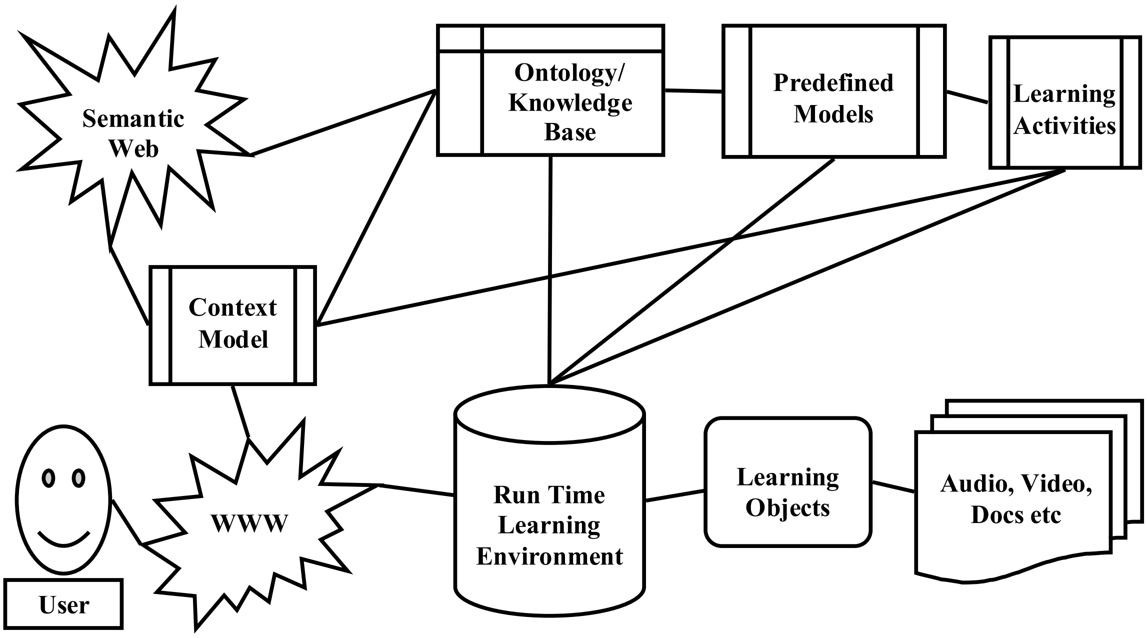

Figure 2. Semantic e-Learning Framework

The semantic e-Learning framework consists of three stages, namely

1. Pre Learning Process
2. Learning Process
3. Post Learning Process

Pre Learning process is designed for Instructors and learners. Instructors indentify the multimedia resources on the web, assigns contextual information to it, design learning path for different types of learners and design learning activities and assessment for individual sessions and whole course. Learners are indentifies and are profiles based on the questionnaire given to it. This profiling is used to identify the personality of the learner based on which a course is catered to it.

Learning process contribute various kinds of learning activities involving, locating learning material, reading material, writing reflection, discussions with peers, self evaluation and revision etc. More precisely this is a learner centric stage where complex learning activity is a combination of simple activities.

Post learning process is the final stage involving reporting and evaluation of learning outcome on both the sides i.e. learner and instructor sides.

Hyunjong Choe et al. [6] have concentrated on the IEEE Learning Technology System Architecture (LTSA) which represents a variety of learning system from different domains. Following **Fig. 3** show the model referred above.

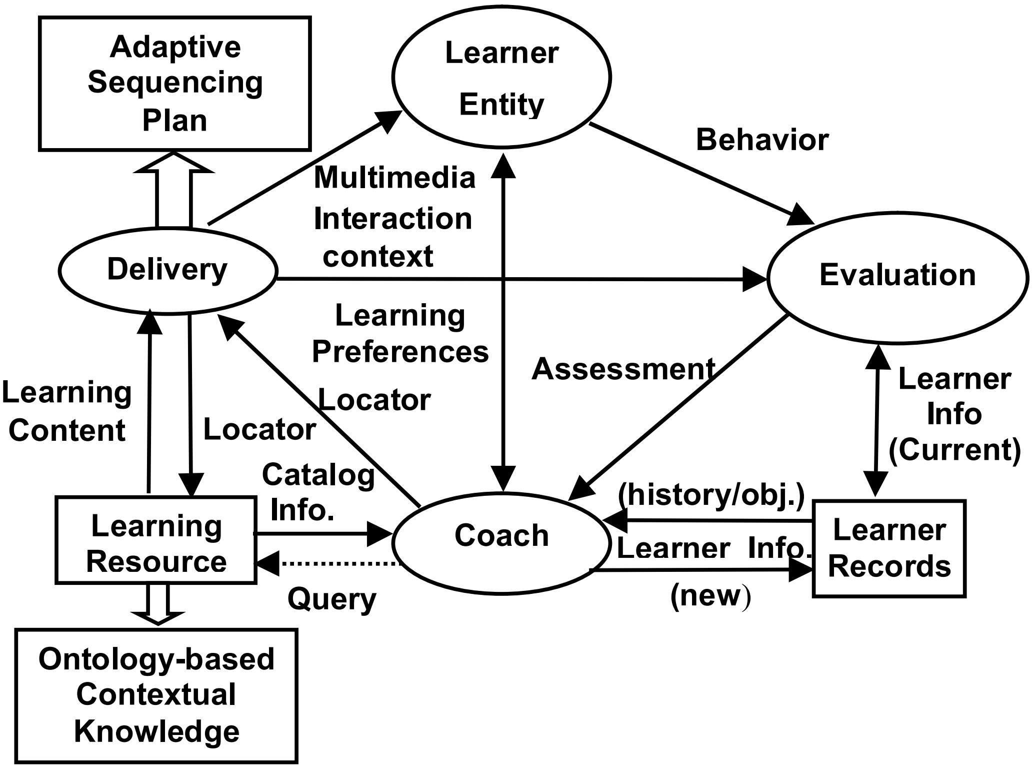

Figure 3. LTSA with Ontological and Contextual Support

In the figure, two squares shows ontology-based model containing adaptive sequencing plan and the ontology-based contextual knowledge. In this model learning resources is combined with ontological knowledge as a resource for contextual learning. Queries are used to search the resulted resources. The evaluation component is used to measure the learner's performance and finally the result is stored in the database named "learner records". Coach component uses learner records to locate a new context.

4 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. [7].

PART Classifiers:

PART works by building a rule and removing its cover, as in the separate-and-conquer technique, repeatedly until all the instances are covered. The rule construction stage differs from standard separate-and-conquer methods because a partial pruned decision tree is built for a set of instances, the leaf with the largest coverage is made into a rule, and the tree is discarded. The pruned decision tree helps to avoid the over pruning problem of methods that immediately prune an individual rule after construction.

It builds a partial decision tree to obtain a rule. It uses C4.5‘s procedures to build a tree. It uses separate-and-conquer strategy. It builds a partial C4.5 decision tree in every iteration and makes the "best" leaf into a rule. PART obtains rules from partial decision trees. It builds the tree using C4.5’s heuristics with the same user-defined parameters as J48. Three rules are found and are intended to be processed in order, the prediction generated for any test instance being the outcome of the first rule that fires. The last, “catch-all” rule will always fire. As with J48, the numbers in parentheses that follow each rule give the number of instances that are covered by the rule followed by the number that are misclassified (if any). PART Class for generating a PART decision list uses separate-and-conquer strategy. Builds a partial C4.5 decision tree in each iteration and makes the ”best” leaf into a rule [8] [9] [10].

FT Classifiers:

FT Classifier for building Functional trees, which are classification trees that could have logistic regression functions at the inner nodes and/or leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.

FT combines a standard univariate decision tree, such as C4.5, with linear functions of the attributes by means of linear regressions. While a univariate decision tree uses simple value tests on single attributes in a node, FT can use linear combinations of different attributes in a node or in a leaf. In the constructive phase a function is built and mapped to new attributes. A model is built using the constructor function. This is done using only the examples that fall at this node. Later, the model is mapped to new attributes. The constructor function should be a classifier or a regresssor depending on the type of the problem. In the former the number of new attributes is equal to the number of classes, in the latter the constructor function is mapped to one new attribute. Each new attribute is computed as the value predicted by the constructed function for each example. In the classification setting, each new attribute value is the probability that the example belongs to one class given by the constructed model. The merit of each new attribute is evaluated using the merit-function of the univariate tree, and in competition with the original attributes [8] [11].

5 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.**

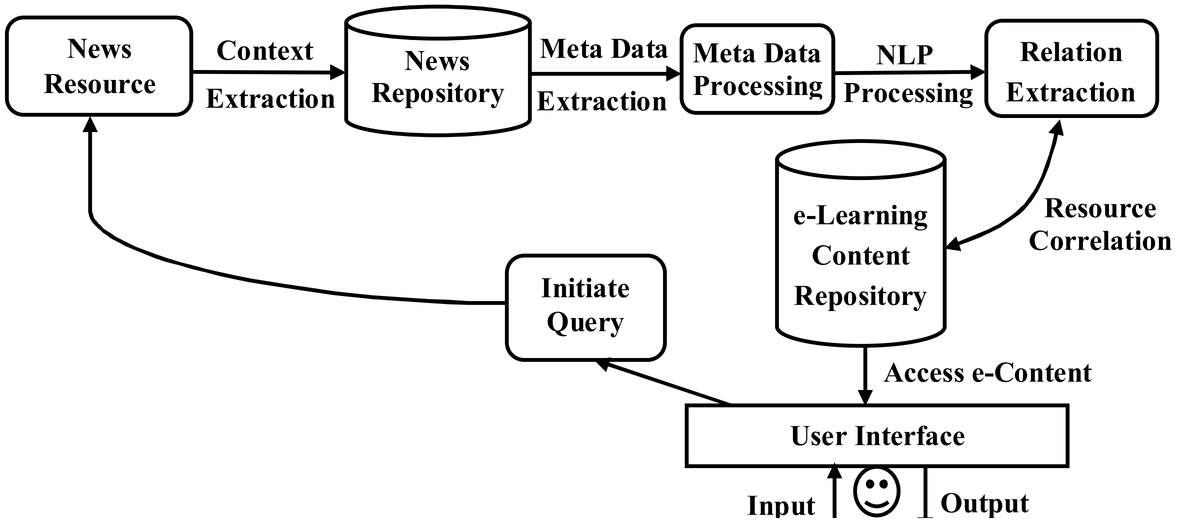

Figure 4. Flow diagram of the model

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 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 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.

6 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.

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**

7 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, two classifier PART and FT 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 and Table 4.

PART works by building a rule and removing its corner, until all instances are covered. It uses C4.5 procedure to build a tree. The performance for Indian News repository has given 96.1% TP and 0.7% FP and area under ROC curve is 99.8%.

FT, a functional true algorithm use linear regression with standard Univariate Decision Tree. A mapping of model to each attribute is done. This makes FT to achieve high accuracy. Using FT 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 and Table 5. The algorithm can be used for classifying the News so that it can then be made available to the e-Learner as per his /her selection.

Table 2. Confusion Matrix For PART

Classified as →	**Education**	**Business**	**Criminal**	**Technology**	**Politics**	**Medical**	**Sports**
Education	**55**	0	0	1	3	0	0
Business	1	**118**	1	1	1	0	1
Criminal	2	0	**78**	0	2	0	0
Technology	2	1	1	**33**	1	1	0
Politics	0	0	0	1	**150**	1	1
Medical	0	0	0	2	0	**44**	0
Sports	0	1	0	0	0	0	**146**

Table 3. Table Showing True Positive and False Positive Rate of PART

Class ↓	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**ROC Area**
Education	93.2%	0.8%	91.7%	93.2%	99.7%
Business	95.9%	0.4%	98.3%	95.9%	99.9%
Criminal	95.1%	0.4%	97.5%	95.1%	99.8%
Technology	84.6%	0.8%	86.8%	84.6%	99.4%
Politics	98%	1.4%	95.5%	98%	99.6%
Medical	95.7%	0.3%	95.7%	95.7%	99.9%
Sports	99.3%	0.4%	98.6%	99.3%	1%
Weighted Avg. →	96.1%	0.7%	96.2%	96.1%	99.8%

Table 4. Confusion Matrix For FT

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 5. Table Showing True Positive and False Positive Rate of FT

Class ↓	TP Rate	FP Rate	Precision	Recall	ROC Area
Education	100%	0%	100%	100%	100%
Business	100%	0%	100%	100%	100%
Criminal	100%	0%	100%	100%	100%
Technology	100%	0%	100%	100%	100%
Politics	100%	0%	100%	100%	100%
Medical	100%	0%	100%	100%	100%
Sports	100%	0%	100%	100%	100%
Weighted Avg. →	100%	0%	100%	100%	100%

A consolidated performance of the two algorithms used to process News data set can be seen below in Table 6.

Table 6. Showing Correct and Wrong Prediction of Classifier.

Classifier →		PART		FT	
News Category	**Actual No. Of News**	**Correct**	**Wrong**	**Correct**	**Wrong**
Business	**123**	**118**	05	**123**	00
Criminal	**82**	**78**	04	**82**	00
Education	**59**	**55**	04	**59**	00
Medical	**46**	**42**	02	**46**	00
Politics	**153**	**150**	03	**153**	00
Sports	**147**	**146**	01	**147**	00
Technology	**39**	**33**	06	**39**	00
Total	**649**	**624**	85	**649**	00
Percentage →		**88.44**	3.85	**100**	00

Here it is to be noted that FT algorithms gives 100% TP rate. Overall PART, and FT algorithms are found suitable for the repository that we have used for training and Testing.

8 CONCLUSIONS

This paper has designed a model which will help the e-Content to refer the latest information in the form of News to get dynamically attached to e-Contents, hence empowering the effectiveness of the e-Learning process by making latest information available to the learner using the framework that we designed.

As per the previous discussion identification of news from dynamic resources can be done with the propose model, we use two classifier i.e. PART and FT to analyze the data sets. As a result it is found that FT algorithm performs well in categorizing all the News. Overall Performance of PART algorithm is acceptable, except 3 News from Education are classified into Politics, 2,2 Criminal News is distributed into Education & Politics, 2 News form Technology are classified into Education and 2 News from Medical category is distributed into Technology. For overall data set detection rate (True Positive rate) for FT is 100% and whereas PART is 96.1%. Hence FT is good classifier as compare to PART classifier. This also can be seen from Table 3 Table 5 and Table 6 above.

9 REFERENCES

[1] Bassoppo-Moyo & Temba.C., “Evaluating e-learning: A front-end, process and post hoc approach.” International Journal of Instructional Media, 33(1). Retrieved October 28, 2007, from ProQuest database.

[2] Weihong Huang & Alain Mille, “ConKMel: a contextual knowledge management framework to support multimedia e-Learning.” Published online: 8 July 2006, Springer Science + Business Media, LLC 2006, pp 205-219

[3] SCORM (2003) “Advanced distributed learning initiative, sharable content object reference model (SCORM).” http://www.adlnet.org/

[4] LOM (1999) “Learning object metadata working group, learning object metadata.” IEEE P1484.12.1-2002 http://ltsc.ieee.org/wg12/

[5] Weihong Huang, David Webster, Dawn Wood and Tanko Ishaya, “An intelligent semantic e-learning framework using context-aware Semantic Web technologies”, British Journal of Educational Technology, Vol 37 No 3, pp 351–373, 2006

[6] Hyunjong Choe, Taeyoung Kim & Chungbuk Korea, "An Enhanced LTSA Model Providing Contextual Knowledge for Intelligent e-Learning Systems", Journal Information Science and Engineering, 2005, pp. 849-858,

[7] http://en.wikipedia.org/wiki/Classification

[8] 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.

[9] R.P.Datta & Sanjib Saha., “An Empirical comparison of rule based classification techniques based classification techniques in medical databases.” 2nd International Congress on Pervasive Computing and Management 12-13, Dec 2009 at Sydney, Australia, pp 1-14

[10] Geoffrey Holmes, Mark Hall and Eibe Frank., “Generating Rule Sets from Model Trees.” Department of Computer Science University of Waikato, New Zealand

[11] Trilok Chand Sharma & Manoj Jain., “WEKA Approach for Comparative Study of Classification Algorithm.” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013, pp 1925-1931