ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES AND DATA MINING

Kanakam Sadhi Kumar*[1], Dr. Arpana Bharani[2] and Dr. Ritesh Yadav[3]
[1,2]Department of Computer Science, Dr. A.P.J. Abdul Kalam University, Indore
[3]Department of Physics, Dr. A.P.J. abdul Kalam University, Indore
[1]sadhi.kanakam@gmail.com

ABSTRACT

Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral. This paper focuses on the several machine learning techniques which are used in analyzing the sentiments and in opinion mining. Sentimental analysis with the blend of machine learning could be useful in predicting the product reviews and consumer attitude towards to newly launched product. This paper presents a detail survey of various machines learning techniques and then compared with their accuracy, advantages and limitations of each technique. On comparing we get 85% of accuracy by using supervised machine learning technique which is higher than that of unsupervised learning techniques. Information extraction is concerned with applying natural language processing to automatically extract the essential details from text documents. A great disadvantage of current approaches is their intrinsic dependence to the application domain and the target language. This has result in order to directly compare the different extraction algorithm for different tasks. In this research author develop a model which is based on Machine Learning for Information extraction using scientific articles. With its astronomical growth over the past decade, the Web becomes huge, diverse and dynamic. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. To extract the query result page by using the combining tag and value similarity methods are used. A large amount of information on the web is presented in regularly structured objects. A list of such objects in a Web page often describes a list of similar items. Web content mining is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. Text mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Keywords: Information Extraction, Deep Learning, Machine Learning, Data Mining, Web Mining, Scientific Articles.

1. INTRODUCTION

Sentimental Analysis is interpreted as determining the notion of people about distinct existence. Nowadays people are used to review the comments and posts on the product which are known as opinion, emotion, feeling, attitude, thoughts or behavior of the user. Sentimental Analysis is a method for identifying the ways in which sentiment is expressed in texts. Sentimental analysis attempts to divine the posture or notion of a keynoter or author, or author against assertive field or an object. There are many claims in sentiment analysis. First is that, a viewpoint which is treated as positive in one case and will be taken as negative in another case. The next claim is that usually people

don't consider their viewpoint in same form. Almost of all reviews incorporate with both positive as well as negative remarks, which can be feasible by interpreting the sentences each at a time. Finding the opinion sites and monitoring them on the web is somewhat difficult. So there will be a need of robotic opinion mining as well as a summarization system.

1.1 MACHINE LEARNING APPROACH

In artificial intelligence, machine learning is one of its subsections which are proceeding with algorithm that let systems to understand. In machine learning technique it uses unsupervised learning, weakly supervised learning and supervised learning.

1.1.1 Supervised Learning

Supervised machine learning technique associate with the use of a marked feature set to retain some classification function and includes learning of function from the experiment along with its input and output. Supervised learning is task of assuming a function labeled trained data set. Training data set includes set of training examples; each and every example consists of couple of an input data as well as expected output.

1.1.2 Weakly-Supervised and Unsupervised Learning.

In practical these supervised methods cannot be always used, because it needs labeled corpora but they are not available all time. Another option for machine learning is weakly-supervised and unsupervised methods which do not require pre-tagged data. Weakly supervised learning consists of large set of unlabeled data and small set of labeled data. Unsupervised method includes learning device for the input

1.2 LEXICON BASED APPROACH

In lexicon based method it supports a lexicon to achieve sentiment classification through weighting and counting sentiment associated words has to be calculated and labeled. To assemble the viewpoint list there are three major methods are considered: dictionary-based method, corpus-based method and the manual opinion approach.

1.3 MACHINE LEARNING APPROACH

Machine Learning techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it supported by three technologies that are now sufficiently mature [7].

1 Massive data collection,

2 Powerful multiprocessor computers and

3 Data mining algorithms.

1.4 MACHINE LEARNING TECHNIQUES

1. **Artificial Neural Networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. **Decision Trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
3. **Genetic Algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

4. **Nearest Neighbor Method**: A technique that classifies each record in a dataset based on a combination of the classes of the k records most similar to it in a historical dataset (where k ³ 1) sometimes called the k-nearest neighbor technique.
5. **Rule Induction**: The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms [9].

1.5 OBJECTIVES

1. To investigate the issues of information extraction from Scientific Articles.
2. To compare the performances of the various Machine-Learning techniques.
3. To develop Information Extraction Model for data using Machine Learning.
4. To Comparative analysis of extract web information using deep learning Method vs. Bayesian networks,
5. To Compare and Analysis of Information extraction use various Algorithms such as Deep learning algorithm, Naive Bayes and Back Propagation Neural Network Algorithm.

1.6 PROBLEM STATEMENT

In the modern area, the websites considered as the one touch source of all kinds of information needed by an individual. The data stored in the web spaces are numerous and one can refer to any kind of information with the help of websites. Recently, information extracted from the web using programmed methods because of the need of information. As the extraction process becomes viral, the websites are becoming sources of redundant information. Duplication becomes a major issue. Thus, a method needed to extract information from the websites by identifying the relevant information. The main problem faced by extractors is that, a single website contains the same content a number of times and possesses other irrelevant information.

2. REVIEW OF LITERATURE

Nicholas Kushmerick (2000) has presented numerous sources of useful information telephone directories, product catalogs, stock quotes, event listings, etc. Recently, many systems have built that automatically gather and manipulate such information on a user's behalf. However, these resources are usually formatted for use by people (e.g., the relevant content is embedded in HTML pages), so extracting their content is difficult. Most systems use customized wrapper procedures to perform this extraction task. Unfortunately, writing wrappers is tedious and error-prone. As an alternative, we advocate wrapper induction, a technique for automatically constructing wrappers. In this article, we describe six wrapper classes, and use a combination of empirical and analytical techniques to evaluate the computational tradeoffs among them. He measured the number of examples and time required to learn wrappers in each class, and compare these results to PAC models of our task and asymptotic complexity analyses of our algorithms. Summarizing our results, he find that most of our wrapper classes are reasonably useful (70% of surveyed sites can be handled in total), yet can rapidly learned (learning usually requires just a handful of examples and a fraction of a CPU second per example).

Bing Liu, et al (2003) described latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three level hierarchical Bayesian model, in which each item of a collection is model as a finite mixture over an underlying set of topics. Each topic is in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

Tak-Lam Wong and Wai Lam (2004) have developed a probabilistic framework for adapting information extraction wrappers with new attribute discovery. Wrapper adaptation aims at automatically adapting a previously

learned wrapper from the source Web site to a new unseen site for information extraction. One unique characteristic of our framework is that it can discover new or previously unseen attributes as well as headers from the new site.

Paul Viola (2005) have proposed, conditional Markov chain models (CMM) have been used to extract information from semi-structured text (one example is the Conditional Random Field). Applications range from finding the author and title in research papers to finding the phone number and street address in a web page.

Georgios Sigletos, et al (2005) have investigated the effectiveness of voting and stacked generalization -also known as stacking- in the context of information extraction (IE). A new stacking framework proposed that accommodates well-known approaches for IE. The key idea is to perform cross-validation on the base-level data set, which consists of text documents annotated with relevant information, in order to create a meta-level data set that consists of feature vectors. A classifier then trained using the new vectors. Therefore, base-level IE systems are combining with a common classifier at the meta-level. Various voting schemes are presenting for comparing against stacking in various IE domains. Well-known IE systems employed at the base level, together with a variety of classifiers at the meta-level. Results show that both voting and stacking work better when relying on probabilistic estimates by the base-level systems. Voting proved to be effective in most domains in the experiments.

Jordi Turmo, Alicia Ageno, and Neus Catal (2006) have described of online textual sources and the potential number of applications of knowledge acquisition from textual data has lead to an increase in Information Extraction (IE) research. Some examples of these applications are the generation of databases from documents, as well as the acquisition of knowledge useful for emerging technologies like question answering, information integration, and others related to text mining.

Geoffrey E. Hinton, et al (2006) have presented how to use complementary priors" to eliminate the explaining away effects that make inference difficult in densely connected belief nets that have many hidden layers. Using complementary priors, he derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory.

Hal Daum and Daniel Marcu (2006) have proposed most basic assumption used in statistical learning theory is that training data and test data drawn from the same underlying distribution. Unfortunately, in many applications, the in-domain" test data is drawn from a distribution that is related, but not identical, to the "out-of-domain" distribution of the training data. He considers the common case in which labeled out-of-domain data is plentiful, but labeled in-domain data is scarce. He introduce a statistical formulation of this problem in terms of a simple mixture model and present an instantiation of this framework to maximum entropy classifiers and their linear chain counterparts.

Tak-Lam Wong and Wai Lam (2007) have developed a novel framework that aims at automatically adapting previously learned information extraction knowledge from a source Web site to a new unseen target site in the same domain. Two kinds of features related to the text fragments from the Web documents investigated.

Wenyuan Dai, Qiang Yang, et al (2007) have discussed traditional machine learning makes a basic assumption: the training and test data should be under the same distribution. However, in many cases, this identical distribution assumption does not hold. The assumption might be violate when a task from one new domain comes, while there only labeled data from a similar old domain. Labeling the new data can be costly and it would be a waste to throw away all the old data.

Chia-Hui Chang, et al (2010) have discussed the Internet presents a huge amount of useful information which is usually formatted for its users, which makes it difficult to extract relevant data from various sources. Therefore, the availability of robust, flexible Information Extraction (IE) systems that transform the Web pages into program-friendly structures such as a relational database will become a great necessity. Although many

approaches for data extraction from Web pages have developed, there has been limited effort to compare such tools.

Aurangzeb Khan, et al (2010) have discussed with the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic categorization of documents became the key method for organizing the information and knowledge discovery. Proper classification of e-documents, online news, blogs, e-mails, and digital libraries need text mining, machine learning, and natural language processing techniques to get meaningful knowledge.

Xiaoyan Ren and Yunxia Fu (2010) have described based on the survey of contemporary web information extraction theory, this paper studies the frequently discussed but insufficiently-solved problem: data extracting from web pages containing several structured records, and proposes a new approach called IEBID Tech (Information Extraction based on Improved Dom Tree) which is mainly composed of three steps. At step 1, the given page initially segmented into several blocks according to html delimiters after the transformation of a DOM tree, and the redundant blocks subsequently removed, which then followed by the induction of extraction rules at step2 and the extraction of structured data at step 3. Large numbers of experiments from diverse domains' web pages show that both recall and precision rates are greater than 90%. That is this approach is able to extract data more accurately.

Jer Lang Hong, Eu-Gene Siew, et al (2010) have studied structured records of web pages and the relevant problems associated with the extraction and alignment of these structured records. Current automatic wrappers are complicated because they take into consideration the problems of locating relevant data region using visual cues and the use of complicated algorithms to check the similarity of data records. In this paper, we develop a non-visual automatic wrapper, which questions the need for complex visual based wrappers in data extraction.

3. PROPOSED METHODOLOGY

3.3.1 Proposed Approach

The proposed approach deals with a web information extraction method through deep learning architecture for scientific article. Deep learning is fairly a new area of machine learning and neural network research. It uses neural networks having several hidden layers for finding hierarchical representations of data, starting from observations towards more and more abstract representations. Traditionally, back propagation type algorithms have used for learning appropriate representations of data in such multilayer networks. However, when the number of hidden layers in a feed forward neural network is larger too, such learning algorithms suffer from several problems, which often make it impossible to find satisfactory representations of the data. Hinton and Salakhutdinov found that unsupervised pre-training of the hidden layers using restricted Boltzmann machines helps to obtain much better representations of the data. The data is not only to learn the nonlinear mapping between input and output vectors, but also a good representation of the input data. The learning results provided by restricted Boltzmann machines can be refined and improved further by using back propagation type supervised learning algorithms. Trained in this manner, deep networks have provided world-record results in many classification and regression benchmark problems, leading to a Renaissance of neural network research. The usual web content extraction methods concentrate only on extracting the content without checking whether it is relevant or not. The proposed approaches compare various Algorithm method and Architecture with the performance evaluations. This work provides web URLs or web document are extract content or non-content.

3.4 METHODOLOGY

3.4.1 Machine Learning Algorithm

Deep learning (deep machine learning, or deep structured learning, or hierarchical learning, or sometimes DL) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations.

3.4.1.1 Machine Learning Architecture

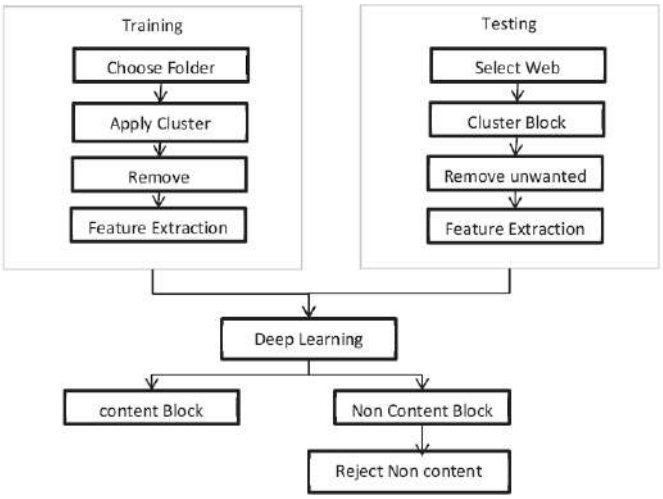

Figure 3.1 Deep Learning Architecture.

- **Cluster Blocks:** Clustering is used to divided the input web document into small parts. The web document contains unwanted elements such as tag, side bar, advertisements, HTML content etc.
- **For example :**

<!DOC TYPE html>

<html>

<body>

<h1>My First Heading</h1>

My first paragraph

</body>

</html>

This is one of the html documents. Then apply clustering the above document divided into many parts such as tag, side bar, comments, and html content.

- **Remove unwanted block:** After cluster, we get many blocks then we will remove the unwanted blocks such as tag, side bar, and comments. Finally, we get only html content, which has given as input of feature extraction.
- **Feature Extraction:** Feature extraction means extracts many features from the html content.

- **Data Unit (DU) and Text Features (TF):** These features are applicable to text nodes, including composite text nodes involving the same set of concepts, and template text nodes.
- **Data Content (DC):** The data units or text nodes with the same concept often share certain keywords.
- **Presentation Style (PS):** This feature describes how a data unit displayed on a webpage. It consists of style features such as font size color etc.
- **Data Type (DT):** Each data unit has its own semantic type although it is just a text string in the HTML code such as Date time, string integer etc.

- **Tag Path (TP):** A tag path of a text node is a sequence of tags traversing from the root of the SRR to the corresponding node in the tag tree. The above steps are apply in testing phase and get the features of html document these features given to the Deep learning algorithm for extraction.

3.4.1.2 Deep Learning Algorithm

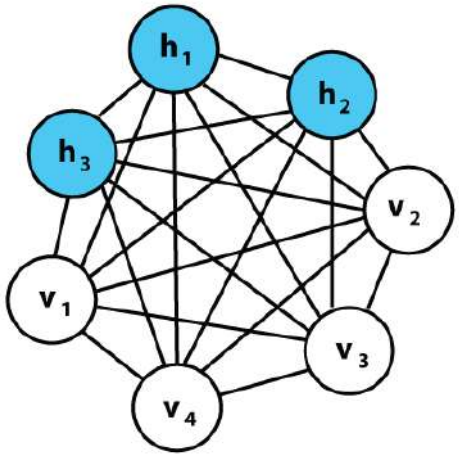

Figure 3.2 Boltzmann Machine

- One input layer and one hidden layer
- Typically binary states for every unit
- Stochastic (vs. deterministic)
- Recurrent (vs. feed-forward)
- Generative model (vs. discriminative): estimate the distribution of observations(say p(image)), while traditional discriminative networks only estimate the labels(say p(label image))
- Defined Energy of the network and Probability of a unit's state (scalar T referred to as the "temperature").

$$E(s) = -\sum_i a_i s_i - \sum_{i<j} s_j w_{i,j} s_i \quad (3.1)$$

$$P(s_j = 1) = \frac{1}{1 + e^{(-\frac{\Delta E_j}{T})}} = \sigma((s_j + \sum_{i=1}^{m} w_{i,j} s_i)/T) \quad (3.2)$$

- A bipartite graph: no interlayer connections, feed-forward
- RBM does not have T factor, the rest are the same as BM
- One important feature of RBM is that the visible units and hidden units are conditionally independent, which will lead to a beautiful result later on:

$$P(v|h) = \prod_{i=1}^{m} P(v_i|h) \quad (3.3)$$

$$P(h|v) = \prod_{j=1}^{n} P(h_j|v) \quad (3.4)$$

- Two characters to define a Restricted Boltzmann Machine:
- States of all the units obtained through probability distribution.
- Weights of the network obtained through training (Contrastive Divergence).
- As mentioned before, the objective of RBM is to estimate the distribution of input data. Moreover, this goal fully determine by the weights, given the input.
- Energy defined for the RBM:

$$E(v,h) = -\sum_i a_i v_i \sum_j b_j h_j - \sum\sum_j h_j w_{i,j} v_i \quad (3.5)$$

- Distribution of visible layer of the RBM (Boltzmann Distribution):

$$P(v) = \frac{1}{z}\sum_h e^{-E(v,h)} \quad (3.6)$$

- Z is the partition function defined as the sum of over all possible configurations of {v, h}
- Training for RBM: Maximum Likelihood learning
- the probability over a vector x with parameter W(weights) is:

$$P(x;W) = 1/Z(W)\, e^{-E(x;W)} \quad (3.7)$$

$$Z(W) = \sum_x e^{-E(x;W)} \quad (3.8)$$

4. EXPERIMENTAL SET UP

4.1 INTRODUCTION

After extracts all algorithm the document is group into content and non-content document. The content block has only the important words from the URL. A noncontact block has unwanted comments tags and advertisements. We extract only the content blocks so remove the non-content block. Finally remove the non-content and get content only the output.

4.2 EXPERIMENTAL SETUP

The experimental results of the proposed method web data extraction web document clustering presented in this section. The proposed approach has implemented in the result shows the output of document classification based on deep learning, naïve Bayes, and back propagation algorithm. The web URL is the input and test dataset for this project and find content block for this web URL. This project removes the html tags, comments, advertisements etc..., implemented as MATLAB 2013a, and the experimentation performed on a 3.0 GHz core i5 PC machine with 4 GB main memory. For experimentation, we have taken many web pages, which contained all the noises such as Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and Other Uninteresting Data. The Web Pages then subjected to process through the proposed deep learning network to web content extraction. The performances of the proposed approach are analyses using performance measures precision, recall, and F-measure. The following results show that. The comparative result is show in below screen short.

4.3 PERFORMANCE EVALUATION

The performance evaluation used to compare existing and proposed system. The performance metrics are precision, recall, and F-Measure. These metrics used to compare the different web pages.

Performance Measures:

1. Precision:

Precision is the percentage of the relevant data records identified from the web page. The proposed approach has selected a set of web documents in the evaluation process. The different blocks evaluated here based on the precision parameter. The precision defines the relevance of the extracted blocks by the proposed Algorithms based web content extraction. The precision is the fraction of retrieved instances that are relevant to the find.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad \text{……………………………………………………(1)}$$

Where,

TP = True Positive (Equivalent with Hits)

FP = False Positive (Equivalent with False Alarm)

2. Recall

The recall is the fraction of relevant instances that retrieved according to the query.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad \text{………………………………………….……(2)}$$

Where,

TP = True Positive (Equivalent with Hits)

FN = False Negative (Equivalent with Miss)

3. F-Measure

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can calculated as,

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \qquad \text{…………………………...………(3)}$$

4.4 PERFORMANCE ANALYSIS

The documents collected from the internet and the web pages collected contain both relevant and irrelevant contents as per the need of the user. The role of the proposed Algorithms theology is to extract the relevant contents from the web pages by identifying them accurately. The proposed approach selected a set of web documents from the internet and manually extracted its blocks based on HTML tag. Then they analyzed and the important blocks identified. After the manual calculations, the proposed deep learning method subjected to process with the extracted web pages. The content extracted by the proposed deep learning method compared with content identified manually for evaluating the accuracy of the proposed deep learning based web content extraction.

4.4.1 Deep Learning Algorithm

The proposed approach has selected a set of web documents in the evaluation process. The different blocks are evaluated here based on the precision, Recall and FMeasures parameter.

Performance analysis based on precision in DLA:

Table 4.1 Precision Value for DLA

Different Web Pages	Precision (%)
http://www.international.ucla.edu/korea/	78
http://www.coronaregional.com/	80
http://www.pinchin.com/newsletter-list	75
http://www.medwebplus.com/	90
http://www.thieme.com/index.php?	88
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	93
http://www.eicar.org/	95

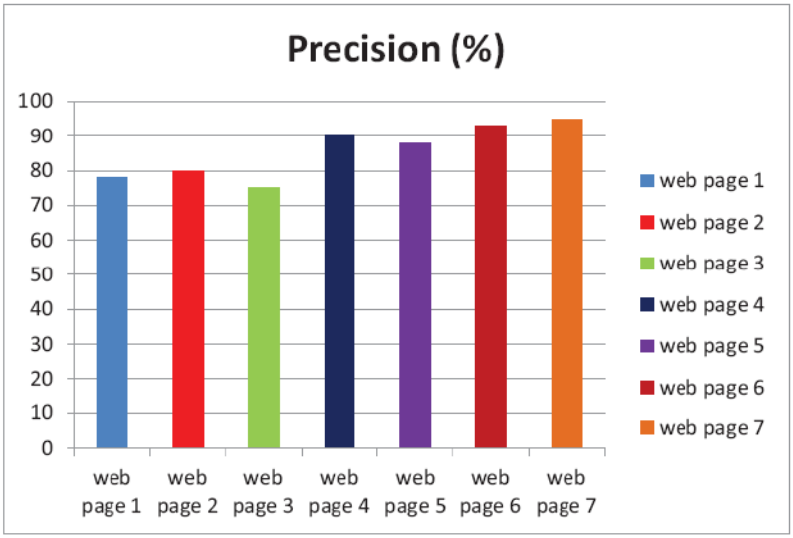

Figure 4.12 Precision value of DLA

The above graph and table shows the result of precision for different web pages using Deep learning. We consider different web pages for comparison. The precision values are same as accuracy. The above definition and formula used to find the precision value.

Performance analysis based on Recall in DLA:

Table 4.2 Recall Value for DLA

Different Web Pages	Recall (%)
http://www.international.ucla.edu/korea/	75
http://www.coronaregional.com/	76
http://www.pinchin.com/newsletter-list	72
http://www.medwebplus.com/	79
http://www.thieme.com/index.php?	77
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	81
http://www.eicar.org/	82

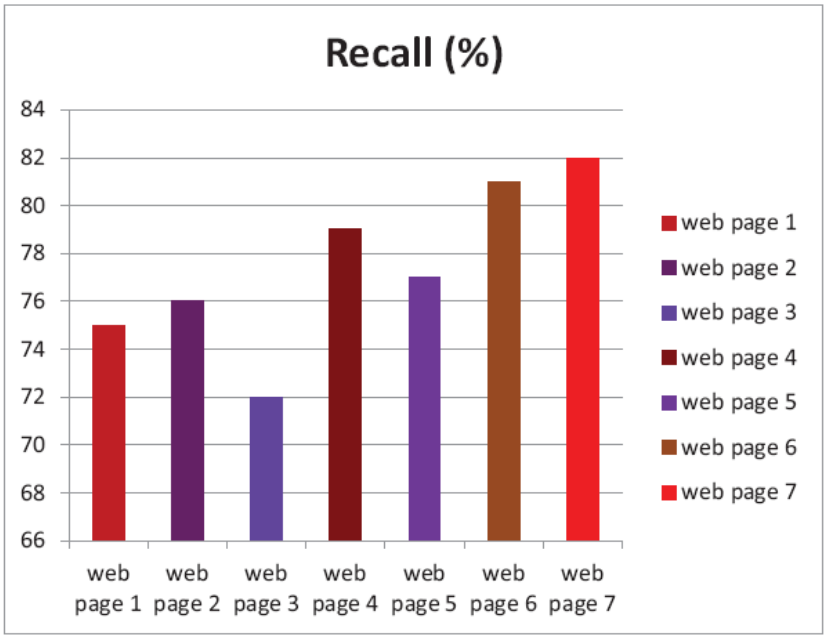

Figure 4.13 Recall value of DLA

The recall value is same as hit rate and sensitivity value. This will calculated by using true positive and false negative values. The above chart shows the result of recall value for all tables using deep learning.

Performance analysis based on F-Measure in DLA:

Table 4.3 F-Measure Value for DLA

Different Web Pages	F-Measure (%)
http://www.international.ucla.edu/korea/	76
http://www.coronaregional.com/	79
http://www.pinchin.com/newsletter-list	75
http://www.medwebplus.com/	74
http://www.thieme.com/index.php?	73
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	82
http://www.eicar.org/	85

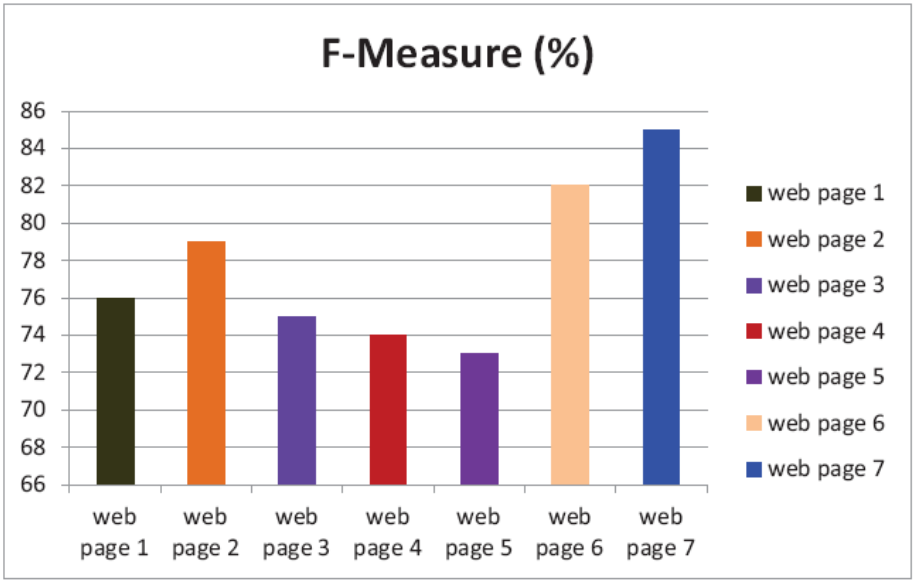

Figure 4.14 F-Measure value of DLA

The above graph and table shows the result of f-measure rate for different web pages using Deep Learning. We consider different web pages for comparison. The above definition and formula used to find the F-measure value.

Performance analysis based on precision, Recall and F-Measure in DLA:

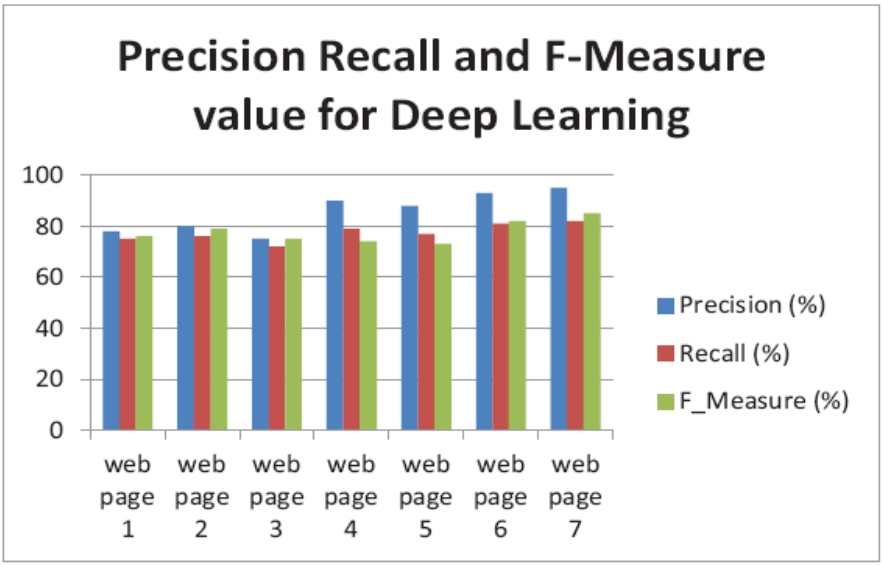

Figure 4.15 Precision, Recall and F-Measure value for Deep Learning

The above chart shows the overall result of deep learning algorithm for all web pages.

4.4.2 Naive Bayesian Algorithm

Naive Bayesian propagation Algorithm Performance analysis based on

Precision, Recall, and F- Measure Values:

Table 4.4 Precision Value for NB

Different Web Pages	Precision (%)
http://www.international.ucla.edu/korea/	73
http://www.coronaregional.com/	75
http://www.pinchin.com/newsletter-list	70
http://www.medwebplus.com/	84
http://www.thieme.com/index.php?	82
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	86
http://www.eicar.org/	83

Naive Bayesian Performance Analysis Based On Precision:

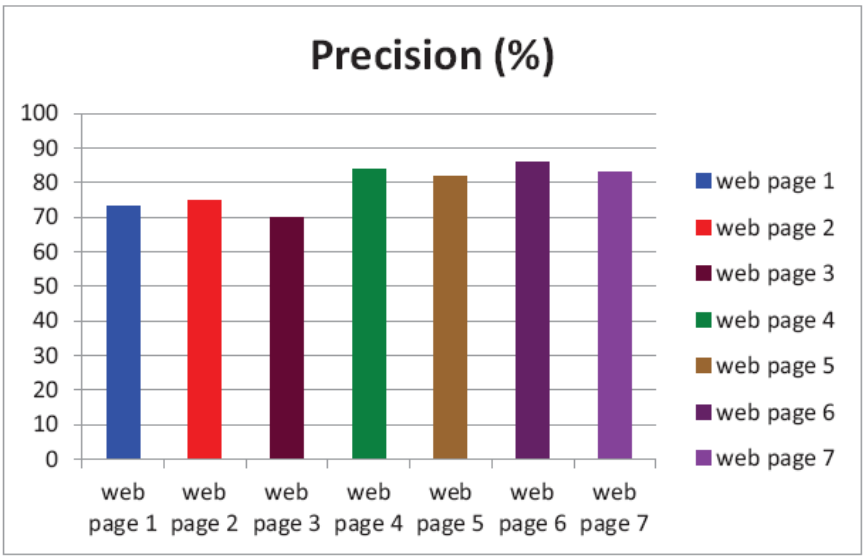

Figure 4.16 Precision value of NB

The precision value is same as accuracy value. This value calculated using true positive and false positive values. True positive means the correct contents in the content block. False positive means wrong contents in the content block.

Naive Bayesian Performance analysis based on Recall:

Table 4.5 Recall Value for NB

Different Web Pages	Recall (%)
http://www.international.ucla.edu/korea/	70
http://www.coronaregional.com/	74
http://www.pinchin.com/newsletter-list	67
http://www.medwebplus.com/	75
http://www.thieme.com/index.php?	72
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	77
http://www.eicar.org/	79

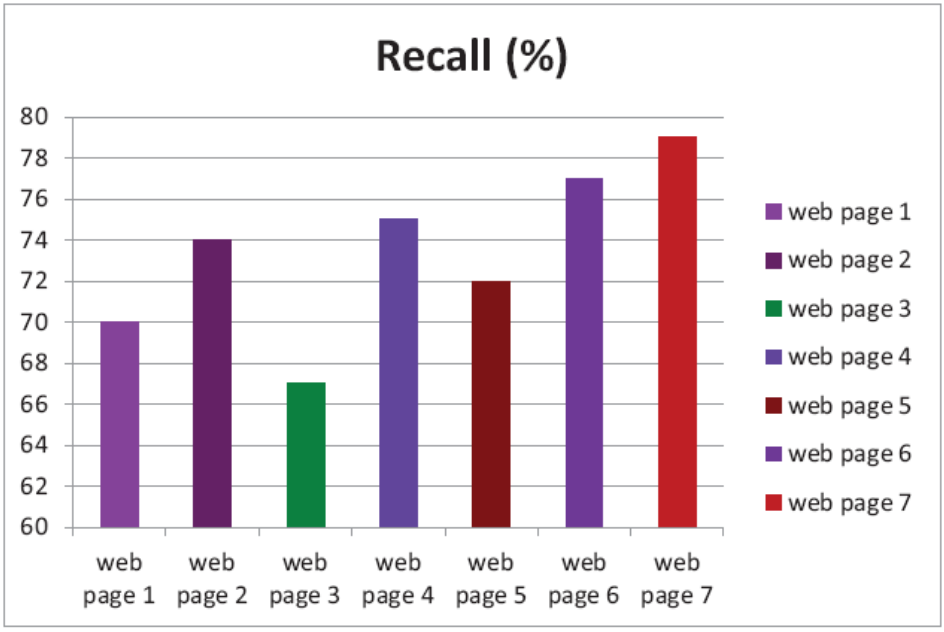

Figure 4.17 Recall value of NB

The above graph and table shows the result of recall rate for different web pages using Naïve Bayesian. We consider different web pages for comparison. The recall values are same as error rate. The above definition and formula used to find the recall value. The recall value will calculated by using true positive and false negative. False negative means the content in the non-content block.

Naive Bayesian Performance analysis based on F-Measure:

Table 4.6 F-Measure Value for NB

Different Web Pages	F_Measure (%)
http://www.international.ucla.edu/korea/	72
http://www.coronaregional.com/	75
http://www.pinchin.com/newsletter-list	70
http://www.medwebplus.com/	69
http://www.thieme.com/index.php?	67
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	77
http://www.eicar.org/	80

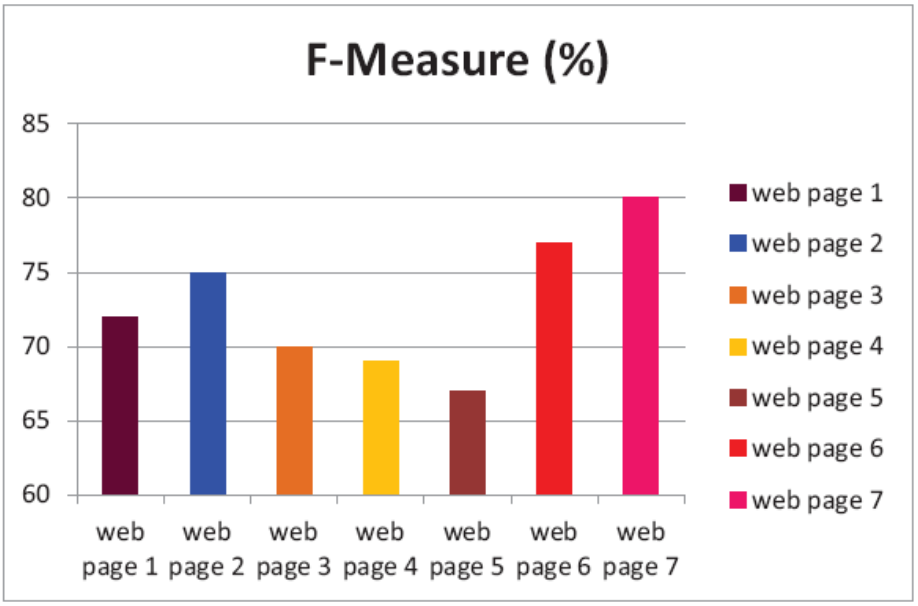

Figure 4.18 F-Measure value of NB

The above graph and table shows the result of f-measure rate for different web pages using Naïve Bayesian. We consider different web pages for comparison. The above definition and formula used to find the F-measure value.

Naive Bayesian Performance analysis based on Precision, Recall and F-Measure:

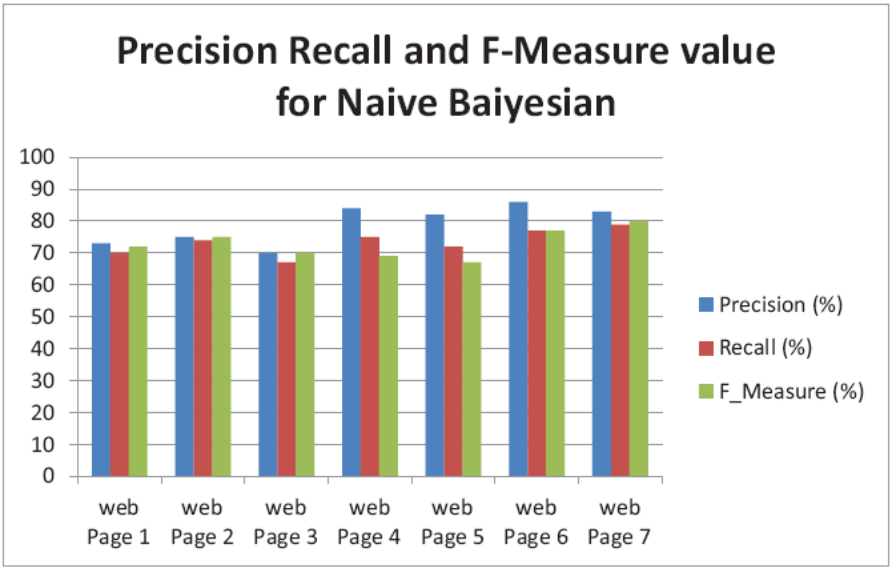

Figure 4.19 Precision Recall and F-Measure value for NB

The above graph shows the overall result Naïve Bayesian classification algorithm. The chart shows comparison 3 metrics for all web pages.

6.4.3 Back Propagation Algorithm

Back propagation Algorithm Performance analysis based on Precision, Recall, and F- Measure Values:

Back propagation Algorithm Performance analysis based on Precision:

Table 4.7 Precision Value for BPNN

Different Web Pages	Precision (%)
http://www.international.ucla.edu/korea/	76
http://www.coronaregional.com/	78
http://www.pinchin.com/newsletter-list	73
http://www.medwebplus.com/	89
http://www.thieme.com/index.php?	85
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	90
http://www.eicar.org/	86

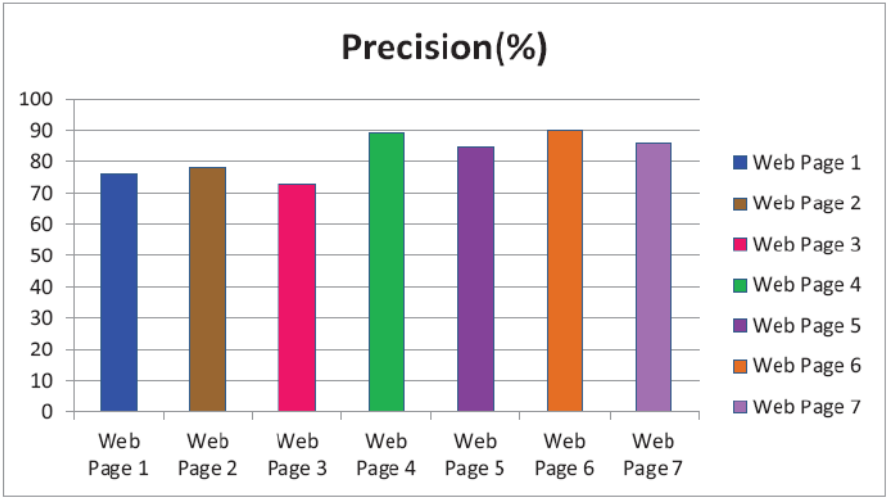

Figure 4.20 Precision value of BPNN

The above graph and table shows the result of precision for different web pages. We consider different web pages for comparison. The precision values are same as accuracy. The above definition and formula used to find the precision value.

4.5 COMPARATIVE ANALYSIS

The comparative analysis portions states the comparison of the proposed approach with an existing web content extraction method. Three-classification algorithms for document extraction i.e. find the content and remove other unwanted content for the web page. In previous we presented the deep learning and Bayesian for this work. Both are provides good performance. These works compare the deep learning, Bayesian and BPNN algorithm. For this comparison, we sued the performance metrics such as precision, recall and F-Measure value. The below table and graph show the overall performance of deep learning, Bayesian and back propagation neural network algorithm. The deep learning is best to compare other algorithms. Thus, it can state that the proposed deep learning algorithm based approach is efficient in extracting the web contents by identifying the important Content. The below table present the overall three algorithms such as Deep Learning vs. Naïve Bayes vs. Back Propagation Neural

Networks Performance analysis based on Precision, Recall and F-Measure Values. These Performance analyses provide the deep learning algorithm perform well to compare other Algorithms.

Over all Three algorithm Performance analysis based on Precision, Recall and F-Measure Values

Table 4.10 Comparative Analysis based on the three algorithms

Various Algorithms	DLA			NB			BPNN		
Different Web Pages	Precision (%)	Recall (%)	F-Measure (%)	Precision (%)	Recall (%)	F-Measure (%)	Precision (%)	Recall (%)	F-Measure (%)
http://www.international.ucla.edu/korea/	78	75	76	73	70	72	76	73	74
http://www.coronaregional.com/	80	76	79	75	74	75	78	75	78
http://www.pinchin.com/newsletter-list	75	72	75	70	67	70	73	70	73
http://www.medwebplus.com/	90	79	74	84	75	69	89	78	71
http://www.thieme.com/index.php?	88	77	73	82	72	67	85	76	70
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-les	93	81	82	86	77	77	90	80	79
http://www.eicar.org/	95	82	85	83	79	80	86	81	83

Over all three-algorithm Performance analysis based on Precision, Recall, and F-Measure:

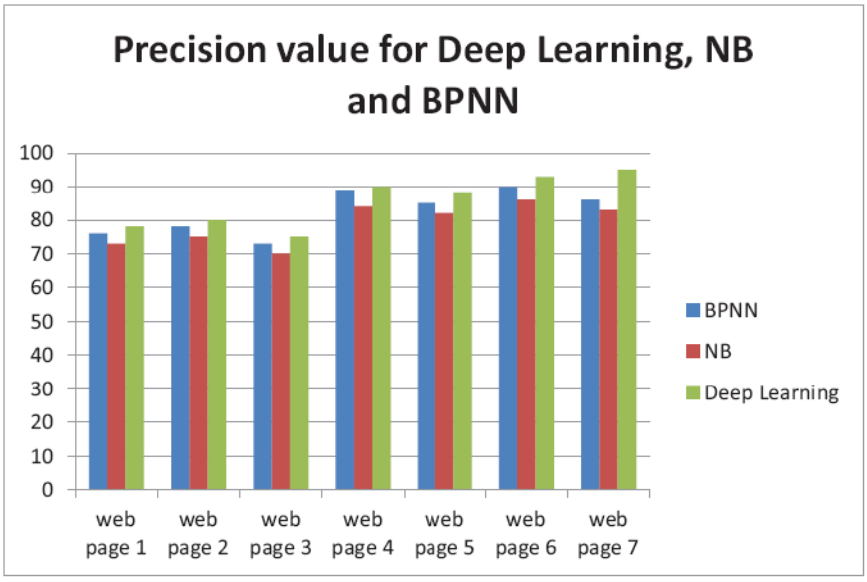

Figure 4.24 Comparative analysis for Precision

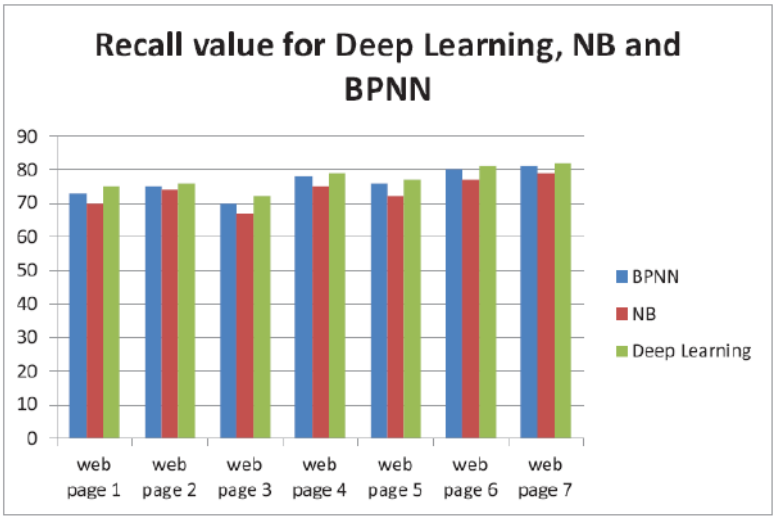

Figure 4.25 Comparative analysis for Recall

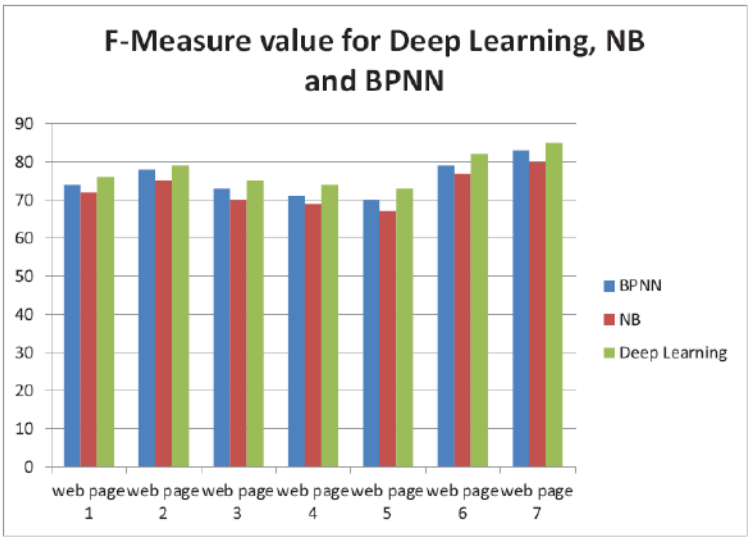

Figure 4.26 Comparative analysis for F- Measure

The above Table & charts shows the result of precision, recall and F-Measure value for all web pages using BPNN, NB and Deep learning classification algorithm.

Average values of Various Methods:

Table 4.11 Average values of Various Methods

Various Method	Precision	Recall	F-Measure
Deep Learning	94	74	73
Bayesian	85	65	63
BPNN	89	72	70

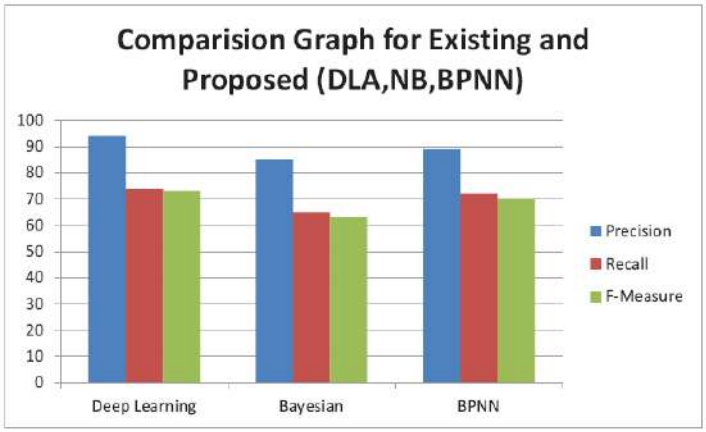

Figure 4.27 Comparison Graph for Existing and Proposed (DLA, NB, BPNN)

The above graph shows the comparative graph for deep learning, Naïve Bayesian and BPNN based on the precision, recall, and F-Measure value. The Deep Learning has the best precision and less error rate values. So compare these three algorithms the deep learning classification algorithm is best for web document extraction.

CONCLUSION

Text mining directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. However, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information. Web text mining is very effective when used in relation to a content database dealing with specific topics. For example, online universities use a library system to recall articles related to their general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines. This allowance of only the most relevant information provided gives a higher quality of results. This increase of productivity is due directly to use of content mining of text and visuals. The main uses for this type of data mining are to gather, categorize, organize, and provide the best possible information available on the WWW to the user requesting the information. This tool is imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information provided to the search engines in order of relevance giving results that are more productive in each search. Web content categorization with a content database is the most important tool to the efficient use of search engines. A customer requesting information on a particular subject or item would otherwise have to search through thousands of results to find the most relevant information to his query. Thousands of results through use of mining text are reduce by this step. This eliminates the frustration and improves the navigation of information on the Web. For the comparative analysis taken the precision, recall, and F-Measure are the performance metrics. The precision rate of Deep Learning gives a high detection rate when compared to Naive Baye's Classifier and BPNN. The recall rate of Naive Baye's is 65% where as, for back propagation is 72% respectively. However, Deep Learning algorithm yields 74% recall rate. Therefore, Deep Learning gives 2% higher recall rate than the BPNN. The sensitivity rate of Deep Learning, which gives a high detection rate when compared Naive Baye's Classifier, and BPNN. The specify rate of Naive Baye's is 63% whereas, for back propagation is 70% respectively. Nevertheless, Deep Learning algorithm yields 72% recall rate. So Deep Learning gives 2% higher specificity rate than the BPNN. As well as the F-Measure, value of Deep Learning which gives a high detection rate when compared to Naive Baye's Classifier and BPNN. Finally, it is examine that Deep Learning is faster and accurate as compared to Naive Baye's and Back propagation. Deep learning systems are possible to implement now because of three reasons: High CPU power, Better Algorithms, and the availability of more data. Over the next few years, these

factors will lead to more applications of Deep learning systems. Deep learning applications are best suited for situations, which involve large amounts of data and complex relationships between different parameters. Solving intuitive problems: Training a Neural network involves repeatedly showing it that: "Given an input, this is the correct output" If this done enough times, a sufficiently trained network will mimic the function you are simulating. It will also ignore inputs that are irrelevant to the solution. Conversely, it will fail to converge on a solution if you leave out critical inputs. Deep learning involves learning through layers, which allows a computer to build a hierarchy of complex concepts out of simpler concepts. Deep learning used to address intuitive applications with high dimensionality. It is an emerging field and over the next few years, due to advances in technology, we are likely to see many more applications in the Deep learning space. I am specifically interested in how

REFERENCES

1. William W. Cohen, and Wei Fan "Learning page-independent heuristics for extracting data from Web pages" TORONTO'99, Pages: 563-574, 1999.
2. Nicholas Kushmerick "Wrapper induction: Efficiency and expressiveness" Artificial Intelligence (Elsevier), 118, Pages: 15–68, 2000.
3. Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo "RoadRunner: Towards Automatic Data Extraction from Large Web Sites" Proceedings of the 27th VLDB Conference, Pages: 1-10, Roma, Italy, 2001.
4. Soumya Ray, and Mark Craven "Representing Sentence Structure in Hidden Markov Models for Information Extraction" Appears in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01), Pages: 1-7, 2001.
5. Bing Liu, Robert Grossman, and Yanhong Zhai "Mining Data Records in Web Pages" ACM Int. Conf. on SIGKDD'03, August 24-27, 2003, Washington, DC, USA.
6. Bing Liu, Robert Grossman, and Yanhong Zhai "Latent Dirichlet Allocation"Journal of Machine Learning Research, 3, Pages: 993-1022, 2003.
7. Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger "Tackling the Poor Assumptions of Naive Bayes Text Classifiers" Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Pages: 1-8, Washington DC, 2003.
8. Oren Etzioni, Craig A. Knoblock, Rattapoom Tuchinda, and Alexander Yates "To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase 170 Price" Int. Conf. on SIGKDD '03, Pages: 2427-2435, August 2003, Washington, DC, USA.
9. Pranam Kolari and Anupam Joshi "webmining: research and practice, IEEE computing in science & engineering, pages: 43-53, july/august 2004.
10. Dragos Arotaritei, and Sushmita Mitra "Web mining: a survey in the fuzzy framework" Fuzzy Sets and Systems (Elsevier) 148, Pages: 5–19, 2004.
11. Valter Crescenzi and Giansalvatore Mecca "Automatic Information Extraction from Large Websites", Journal of the ACM, Vol. 51, No. 5, pp. 731–779. September 2004.
12. Tak-Lam Wong and Wai Lam "A Probabilistic Approach for Adapting Information Extraction Wrappers and Discovering New Attributes" Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04) Pages: 1-8, 2004.
13. Kyung-Joong Kim, Sung-Bae Cho "Fuzzy integration of structure adaptive SOMs for web content mining" Fuzzy Sets and Systems 148, Pages: 43–60, 2004.

14. Michael Azmy "Web Content Mining Research: A Survey" DRAFT Version 1, Pages: 1-15, November, 2005.
15. Paul Viola "Learning to Extract Information from Semi structured Text using aDiscriminative Context Free Grammar" Draft submitted to the conference ACM-SIGIR, Pages: 1-8, 2005.
16. Georgios Sigletos, Michalis Hatzopoulos, Georgios Paliouras and Constantine D. Spyropoulos "Combining Information Extraction Systems Using Voting and Stacked Generalization" Journal of Machine Learning Research, 6, Pages: 1751-1782, 2005.
17. Geoffrey E. Hinton, Simon Osindero, and Yee-Whye The "A fast learning algorithm for deep belief nets" Neural Computation, Pages: 1-16, 2006.
18. Qiang Yang, and Xindong WU "10 Challenging Problems in data mining research" International Journal of Information Technology & Decision Making, Vol. 5, No. 4, Pages: 597–604, 2006.
19. Ioan Pop "An approach of the Naive Bayes classifier for the document classification" General Mathematics Vol. 14, No. 4, Pages: 135–138, 2006.
20. Jordi Turmo, Alicia Ageno, and Neus Catal "Adaptive Information Extraction" ACM Computing Surveys, Vol. 38, No. 2, Article 4, Pages: 1-47, 2006.
21. Geoffrey E. Hinton, Simon Osindero and Yee-Whye The "A Fast Learning Algorithm for Deep Belief Nets" Neural Computation (Elsevier) 18, Pages: 1527–1554, 2006.
22. Hal Daum and Daniel Marcu "Domain Adaptation for Statistical Classifiers" Journal of Artificial Intelligence Research 26, Pages: 101-126, 2006.
23. A.Jebaraj Ratnakumar "An implementation of web personalization using web mining techniques" Journal of Theoretical and Applied Information Technology, 2005-2010 JATIT, Pages: 67-73, 2006.
24. Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Hsiao-Wuen Hon "Webpage Understanding: an Integrated Approach" ACM Int. Conf on KDD'07, Pages: 1-10, August 12–15, 2007, San Jose, California, USA.
25. Tak-lam Wong and Wai lam "Adapting Web Information Extraction Knowledge via Mining Site-Invariant and Site-Dependent Features" ACM Transactions on Internet Technology, Vol. 7, No. 1, Article 6, Pages:1-40, 2007.
26. Wenyuan Dai , Qiang Yang, Gui-Rong Xue, and Yong Yu "Boosting for Transfer Learning" Proceeding of 24th International Conference on Machine Learning, Corvallis, Pages: 1-8, 2007.
27. Insung Jung, and Gi-Nam Wang "Pattern Classification of Back-Propagation Algorithm Using Exclusive Connecting Network" World Academy of Science, Engineering and Technology, 36, Pages: 189-193, 2007.
28. Jen-Yuan Yeh, Hao-Ren Ke, and Wei-Pang Yang "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network" Expert Systems with Applications (Elsevier) 35, Pages: 1451-1462, 2008.
29. Jason Weston, Frederic Ratle, and Ronan Collobert "Deep Learning via Semi- Supervised Embedding", Proceedings of the 25th International Conference on Machine Learning, Pages: 1-8,Helsinki, Finland, 2008.
30. Mohammed Salem Binwahlan, Naomie Salim, and Ladda Suanmali "Swarm Based Features Selection for Text Summarization" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.1, Pages: 175-179, January 2009.