ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0156

A COMPARATIVE ANALYSIS OF WEB INFORMATION EXTRACTION TECHNIQUES DEEP LEARNING vs. NAIVE BAYES vs. BACK PROPAGATION NEURAL NETWORKS IN WEB DOCUMENT EXTRACTION

J. Sharmila[1] and A. Subramani[2]
[1]*Manonmaniam Sundaranar University, India*
E-mail: sharmi_try@yahoo.co.in
[2]*Department of Computer Science, Government Arts College, Dharmapuri, India*
E-mail: subramani.appavu@gmail.com

Abstract

Web mining related exploration is getting the chance to be more essential these days in view of the reason that a lot of information is overseen through the web. Web utilization is expanding in an uncontrolled way. A particular framework is required for controlling such extensive measure of information in the web space. Web mining is ordered into three noteworthy divisions: Web content mining, web usage mining and web structure mining. Tak-Lam Wong has proposed a web content mining methodology in the exploration with the aid of Bayesian Networks (BN). In their methodology, they were learning on separating the web data and characteristic revelation in view of the Bayesian approach. Roused from their investigation, we mean to propose a web content mining methodology, in view of a Deep Learning Algorithm. The Deep Learning Algorithm gives the interest over BN on the basis that BN is not considered in any learning architecture planning like to propose system. The main objective of this investigation is web document extraction utilizing different grouping algorithm and investigation. This work extricates the data from the web URL. This work shows three classification algorithms, Deep Learning Algorithm, Bayesian Algorithm and BPNN Algorithm. Deep Learning is a capable arrangement of strategies for learning in neural system which is connected like computer vision, speech recognition, and natural language processing and biometrics framework. Deep Learning is one of the simple classification technique and which is utilized for subset of extensive field furthermore Deep Learning has less time for classification. Naive Bayes classifiers are a group of basic probabilistic classifiers in view of applying Bayes hypothesis with concrete independence assumptions between the features. At that point the BPNN algorithm is utilized for classification. Initially training and testing dataset contains more URL. We extract the content presently from the dataset. The Three classification algorithm is utilized for the document extraction. The performance evaluation analyses the accuracy, review and F-measure values. The methodology gives a similar investigation of three algorithms with the performance evaluation for Deep Learning, Bayesian and BPNN Algorithm. There are considerable measures of methodologies that have been created in the zone of Web Information Extraction (IE), which concerns how to collect valuable data for further investigation from web pages.

Keywords:

Information Extraction, Back Propagation Algorithm, Neural Network Algorithm, Deep Learning Algorithm

1. INTRODUCTION

The amount of Web data has been rising quickly, primarily with the advancement of Web 2.0 environment, where the clients are urged to give rich content. A lot of Web information is exhibited as a Web document, which occurs in both detail and list. Extraction of net data could be an important method for data integration; but varied net .Web pages may give the same or analogous information utilizing entirely diverse formats or linguistic uses, which makes the addition of information a fascinating task. The Deep Web is the content on the web not available by a search on general search engines, which is likewise called as concealed Web or undetectable Web. The Deep Web contents are accessed by queries submitted to net databases and also the retrieved data, i.e., question results is encased in sites. Web pages are data records. The distinctive Web pages are made progressively and are hard to list by routine crawler based web crawlers, in particular Google and Yahoo. In this paper, we depict this kind of exceptional Web pages as deep Web pages. A noteworthy issue of online web crawlers is that the unit results are a whole Web document. Human exertion is obliged to inspect each of the returned sections to separate exact information. Automatic information extraction frameworks can automate the task of successfully recognizing the pertinent content sections inside of the document. Information Extraction (IE) is concerned with extracting pertinent information from a gathering of archives. It includes methods and algorithms extract knowledge from distinctive data repositories such as transactional databases, data warehouses, text files, WWW and then forwards. Most of the online resources square measure as machine-readable text Mark-up Language (HTML) documents, that square measure seen by net browsers. In this manner, the need for automated, adaptable Web Information Extraction (IE) tools that that extract information and data from the online pages and transfer into a significance and valuable structures for more investigation will turn into an extraordinary need. Classification is a data mining technique used to predict group membership for data instances. A standout amongst the most vital ways that an organization can guarantee the accomplishment of data classification exertion is to keep things vital. The more unpredictable a data classification scheme, the more probable that clients will get to be confounded by it and the exertion will eventually fail. The most well-known way that groups over-specialist classification schemes to create frameworks that basically have an excess of classification levels. Clients then get to be befuddled about the distinctive levels, the types of information that falls into every level and the level of control obliged when storing, processing or transmitting information at diverse levels. Back propagation is a neural network learning algorithm. The field of neural systems was initially aroused by therapists and neurobiologists who tried to create and test computational analogues of neurons. Back propagation learns by iteratively processing a data set of coaching tuples scrutiny the network’s prediction for every tuple with the particular well-known specialise in esteem's net document is similar in concept to a web page. Each Web document has its individual URI. Note that a Web document is not the same as a file: a single Web document can be accessible in a wide range of configurations and dialects, and a single file, for instance a PHP script, may be in charge of producing a substantial number of Web documents with different URL. A Web document is characterized as something that has a URI and can return representations of the identified resource into HTTP requests. In technical literature the term Information Resource is utilized rather than Web document. In recent, the online sites square measure thought of because the one bit supply of a large varies of information required by Associate in nursing individual. The data stored in the web spaces are various and one can refer to any kind of data with the help of web sites. Recently, information is extracted from the web utilizing programmed methods because of the need of information. As the extraction process becomes to be viral, the web sites are getting sources of redundant information. Duplication turns into a noteworthy issue. In this manner, a system is expected extract information from the web sites by identifying the relevant information. The fundamental issue confronted by extractors is that, a single web site includes the same content in numerous times and has other unessential data moreover. In such manner, Tak-Lam Wong and Wai Lam have proposed a web content mining methodology in the exploration with the support of Bayesian networks. The researchers have done learning on separating the web information and attribute disclosure based on the Bayesian approach. Motivated from the exploration, a method is proposed for web content mining methodology based on a deep learning algorithm. The deep learning algorithm provides the benefit in excess of Bayesian networks since Bayesian network is not considered in any learning architecture like to propose this technique. The deep learning methodology serves to identify the relevant content from the web sites through the layer by layer methodology of the deep learning architecture. A web document is similar in concept to a web page. Every web document has its individual URI. Note that a Web document is not the same as a file: a single web document can be accessible in various arrangements and dialects, and a single document, for instance a PHP script, may be in charge of creating a substantial number of web documents with different URIs. A Web document is characterized as something that has a URI and can return representations of the identified asset in response of HTTP requests. In technical literature the term data Resource is used instead of net document. The document clump (or text clustering) is that the application of cluster analysis to matter documents. It's applications in automatic document, organization, topic extraction and quick information retrieval or filter. It includes the utilization of descriptor extraction. Descriptors are sets of words that depict the contents inside the cluster. Document clustering is generally considered to be a centralized process. For instance, document clustering include web document clustering for search users. The utilization of document clustering can be categorized into online and offline. When compared with offline applications, the online applications are typically compelled by efficiency problems. This Proposed methodology provides a Comparative Analysis of three Algorithms such as Deep Learning Algorithm, Naive Bayes and Back Propagation Neural Network Algorithm. In this method a comparative analysis of three algorithms with the performance evaluations of precision, recall and F-Measure can be computed. Finally we conclude that deep learning algorithm is the best method for document extraction.

2. RELATED WORK

Wenyuan Dai et al. [1] has proposed text classification using Naive Bayesian. This paper says to classify text documents across different distributions. The labelled training data are available but it has different distribution from the unlabelled test data. We have developed a transfer-learning algorithm based on the Naive Bayes classifiers, called NBTC. The NBTC algorithm applies the EM algorithm to adapt the NB model learned from the old data for the new. It first estimates the model based under the distribution ($D\ell$) of the training data. Then, an EM algorithm is designed under the distribution (Du) of the test data. KL divergence measures are used to represent distribution distance between the training and test data. An empirical fitting function based on KL divergence is used to estimate the trade-off parameters in the EM algorithm.

Lawrence McAfee [2] has proposed document classification using DBN. This paper states the implementation and testing of a Deep Belief Network (DBN) of document classification. It is important because of the increasing need for document organization, particularly journal articles and online information. Many popular uses of document classification are for email spam filtering, topic-specific search engine or sentiment detection.

Ladda Suanmali [3] has proposed automatic text summarization based on fuzzy logic. It includes title, feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data. The results show that it generates better average precision, recall and f-score to summaries produced by fuzzy method. In Future implement the multi document summarization and fuzzy logic learning methods use of large data set.

Tak-Lam Wong and Wai Lam [4] developed a new attribute discovery via Bayesian learning approach which can automatically adapt the information extraction patterns learned previously in a source web site to new unseen web sites and discover new attributes together with semantic labels. Extensive experiments from more than 30 real time web sites are in three different domains were conducted and the results exhibit that the framework achieves a very promising performance.

Rajendra Kumar Roul [5] has proposed web document clustering using data mining. This paper studies some clustering methods relevant to the clustering document collections and, in consequence, web data. This method of cluster analysis seems to be relevant in approaching the cluster web data. The graph clustering is also described in its methods to contribute significantly in clustering web data. Based on previously presented information, the core section provides an overview approaches to clustering in the web environment.

Jiang Su et al [6] proposed data classification using semi-supervised multi-modal Naive Bayes. It presents Semi-supervised Frequency Estimate (SFE), a novel semi-supervised parameter learning method for MNB. They first point out that EM's objective function, Maximizing Marginal Log Likelihood (MLL), is quite different from the goal of classification learning, i.e. maximizing conditional log likelihood (CLL). Then propose SFE that uses the estimates of word probabilities obtained from unlabelled data, and class conditional probability given a word, learned from labelled data, to learn parameters of an MNB model.

Amit Ganatra [7] has proposed initial classification through back propagation algorithm. This paper says initial classification using genetic and neural network algorithm. The goal of this hybrid algorithm is to perform weight adjustment in order to minimize the Mean Square Error between obtained output and desired output. It is a better way to apply back propagation algorithm first, so that the search space of Genetic algorithm will be reduced. Hence one can overcome the problem of local minima. The proposed algorithm exploits the optimization advantages of GA for the purpose of accelerating neural network training. BP algorithm is sensitive to initial parameters and GA is not. BP algorithm has high convergence speed and while GA is having slow convergence.

Yan Liu [8] proposed a novel deep learning model for query-oriented multidocuments summarization. Accordingly, the empirical validation on three standard datasets, the results not only show the distinguishing extraction ability of QODE but also clearly demonstrate our intention to provide human-like multi document summarization for nature language processing.

Saduf, Mohd Arif Wani [9] proposed the comparative study of learning in neural network. In [9] a new meta-heuristic search algorithm, called cuckoo search (CS), based on cuckoo bird’s behaviour to train BP in achieving fast convergence rate and to avoid local minima problem. Cuckoo search is initializing and passes the best weights to BPNN. Then load training data, initialize all cuckoo nests and pass the cuckoo nests as weights to network. Feed forward neural network runs using the weights initialized with CS to calculate the error backward and CS keeps on calculating the best possible weight at each epoch until the network is converge. The performance of the proposed CSBP algorithm is compared with the ABC-LM, ABC-BP and BPNN algorithms by means of simulation on three datasets such as 2-bit, 3-bit XOR and 4-bit OR. Finally, the simulation results showed that the computational efficiency of BP training process is highly enhanced when coupled with the proposed hybrid method.

Citra Ramadhenal [10] has proposed classification based on error rate. Hybrid algorithms are used to perform weight adjustment in order to minimize the Mean Square Error. First create a model by running the algorithm on the training data. Then test the model to identify a class label for the incoming new data. The classification model using the available training set which needs to be normalized. Then this data is given to the Back propagation algorithm for classification. After applying Back propagation algorithm, genetic algorithm is applied for weight adjustment. The developed model can then be applied to classify the unknown tuples from the given database and this information may be used by decision maker to make useful decision. But it provides less efficiency.

Daniel Soudry [11] proposed data classification in neural network using discrete continuous weight. In [11] intrusion detection and classification using back propagation neural network approach were followed. It first collects the data set then the data is pre-processed. BPN classifier is built for detection and classification of events. In BPN classifier, first design network and set parameters then initialize weights with random values, finally calculate the actual output from the input. Finally, the Results showed are, it classifies instances into several attack types with low detection rate.

3. METHODS

The proposed approach deals with a web content extraction method through deep learning architecture. Deep learning is a fairly new space of machine learning and neural network analysis. It utilizes neural networks having several hidden layers for locating ranked representations of knowledge, starting from observations towards more and more abstract representations. Traditionally, a back propagation algorithm is used for learning appropriate representations of data in such multilayer networks. However, once the quantity of hidden layers in a feed forward neural network is larger than other learning algorithms it encounter a few related issues which often make it impossible to discover adequate Representations of the data. Hinton and Salakhutdinov discovered that unsupervised pre-training of the hidden layers using restricted Boltzmann machines helps to obtain significantly better representations of the data. The data shouldn't be only to be learning the nonlinear mapping between input and output vectors but in addition an excellent representation of the input data. The learning results provided by Boltzmann machines can be refined and improved with ease of back propagation algorithms. By trained on this manner, deep networks have provided results in classification and regression benchmark problems, leading to a revival of neural network research. The usual web content extraction methods concentrate only on extracting the content without checking whether the content is relevant or not. The proposed approaches include diverse algorithm in a comparative manner to evaluate the performance measures of the web contents in an efficient manner. For extraction purpose it offers web URL or web documents.

4. DEEP LEARNING (DL) ALGORITHM

Deep learning (deep machine learning, or deep structured learning, or stratified learning, or generally DL) may be a branch of machine learning supported a group of algorithms that decide to model high-level abstractions in knowledge by victimization model architectures, with complicated structures or otherwise, composed of multiple non-linear transformations. Deep learning is an element of a broader family of machine learning ways supported learning representations of information. associate observation (e.g., associate image) are often diagrammatical in many ways like a vector of intensity values per component, or in an exceedingly additional abstract manner as a group of edges, regions of specific form, etc.. Some representations make it easier to learn tasks (e.g., face recognition or face expression recognition) from examples. One among the guarantees of deep learning is replacement handcrafted options with economical algorithms for unsupervised or semi-supervised feature learning and stratified feature extraction.

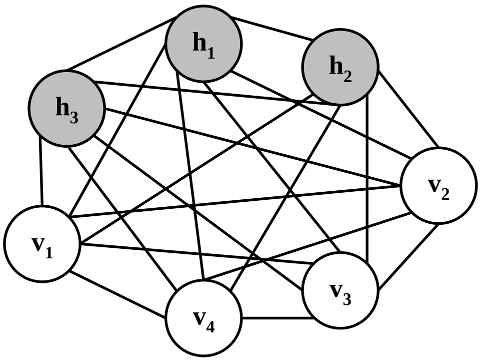

Fig.1. Boltzmann Machine

The Boltzmann Machine modelled with one input layer and one hidden layer generally binary states for each unit. It is processed as stochastic (deterministic), recurrent (feed-forward). A Generative model it estimate the distribution of observations for traditional discriminative networks with labels. The Energy of the network and Probability of a unit's state (scalar *T* is stated as temperature) and it is outlined as follows,

$$E(s) = -\sum_i a_i s_i - \sum_{i<j} s_j w_{i,j} s_i \tag{1}$$

- 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:

$$P(s_j = 1) = \frac{1}{1 + e^{\left(-\frac{\Delta E_j}{T}\right)}} = \sigma\left(\left(s + \sum_{i=1}^{m} w_{i,j} s_i\right) \Big/ T\right) \tag{2}$$

$$P(v|h) = \prod_{i=1}^{m} P(v_i|h) \tag{3}$$

$$P(h|v) = \prod_{j=1}^{n} P(h_j|v) \tag{4}$$

Two characters to define a Restricted Boltzmann Machine are:

- 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. And this goal is fully determined by the weights, given the input.
- Energy defined for the RBM:

$$E(v,h) = -\sum_i a_i v_i \sum_j b_j h_j - \sum\sum_j h_j w_{i,j} v_i \tag{5}$$

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

$$P(v) = \frac{1}{Z} \sum_h e^{-E(v,h)} \tag{6}$$

- *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:

$$P(x;W) = 1/Z(W) e^{-E(x;W)} \tag{7}$$

$$Z(W) = \sum_x e^{-E(x;W)} \tag{8}$$

5. NAIVE BAYESIAN (NB) ALGORITHM:

Bayesian classification provides practical learning algorithms and prior knowledge and determined data are often combined. It provides a helpful perspective for understanding and evaluating several learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input file. Naive Bayes is a conditional probability model for given a problem instance to be classified depicted by a vector *X* = (*x*_1…..*x*_*n*) with *n* instance.

$$P(C_k|x_1 \ldots x_n) \tag{9}$$

The problem with the above formulation is that if n is massive or it will take an oversized range of values, then probability is impossible. We have a tendency to consequently formulate the model to form it additional tractable use of Bayes' theorem, the conditional probability as

$$p(C_k|X) = p(C_k)p(C_k|X)/p(X) \tag{10}$$

Bayesian probability terminology for the above equation can be written as,

Posterior = Prior likelihood/evidence

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on *C* and the values of the features F_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C_k, x_1, \ldots, x_n) \tag{11}$$

This can be rewritten as follows using the chain rule for repeated applications of the definition of conditional probability as:

$$p(C_k, x_1, \ldots, x_n) = p(C_k)p(x_1, \ldots, x_n|C_k) \tag{12}$$

Recently the "Naive" conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$, given the category *C*. This means that

$$p(x_i|C_k, x_j) = p(x_i|C_k) \tag{13}$$

$$p(x_i|C_k, x_j, x_k) = p(x_i|C_k) = p(x_i|C_k) \tag{14}$$

$$p(x_i|C_k, x_j, x_k, x_i) = p(x_i|C_k) \tag{15}$$

For $i \neq j, k$, l Thus, the joint model can be expressed as,

$$p(C_{k,i}|x_1, \ldots, x_j) \propto p(C_k, x_1, \ldots, x_n) \propto p(C_k)p(x_i|C_k)p(x_2|C_k)p(x_3|C_k) \propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{16}$$

This means that under the above independence assumptions, the conditional distribution over the class variable *C* is:

$$p\left(C_{k,i}\middle|x_i,\ldots,x_j\right)={}^{1}\!/_{Z}\,p(C_k)\prod_{i=1}^{n}p\left(x_i\middle|C_k\right) \tag{17}$$

where, the evidence $Z = p(x)$ is a scaling factor dependent only on $x_1,\ldots,x_n$, that is, a constant if the values of the feature variables are known.

6. BACK PROPAGATION NEURAL NETWORKS ALGORITHM

The BPNN algorithm contains three layers appreciate input, output and hidden layer. The BPNN algorithm is employed to compute the errors of the output layer to seek out the errors within the hidden layers. The gradient descent methodology is utilized to calculate the weights and adjustments are created to the network to reduce the output error. The BPNN algorithm has become the standard algorithm used for training multilayer perception. Initially it will find out the errors between the actual and the desired outputs.

$$E_p=\sum_{i=1}^{j}\left(e_i\right)^2 \tag{18}$$

where, e_i is a nonlinear error signal. P denotes in the p^{th} pattern; j is the number of the output units. The gradient descent method is given by,

$$w_{k,i}=\mu\frac{\partial E_p}{\partial w_{k,i}} \tag{19}$$

The Back Propagation calculates errors in the output layer ∂_l, and the hidden layer, ∂_j are using the formulas

$$\partial_l=\mu\left(d_i-y_i\right)f'\left(y_i\right) \tag{20}$$

$$\partial_j=\mu\sum_i\partial_1 w_{l,j}\,f'\left(y_i\right) \tag{21}$$

The back propagation error is used to update the weights and biases in both the output and hidden layers. The weights, $w_{i,j}$ and biases, b_i, are then adjusted using the following formulae:

$$w_{i,j}(k+1)=w_{i,j}(k)+\mu\partial_j y_i \tag{22}$$

$$w_{l,j}(k+1)=w_{l,j}(k)+\mu\partial_j x_l \tag{23}$$

$$b_j(k+1)=b_i(k)+\mu\partial_j \tag{24}$$

where, k is the number of the epoch and μ is the learning rate.

7. EVALUATION SCHEME

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

7.1 PRECISION

Precision is that the proportion of the relevant data records known from the web page. The proposed approach has selected a collection of web documents within the evaluation method. The various blocks are evaluated based on the precision parameter. The precision defines the relevance of the extracted blocks by the proposed web content extraction algorithm. The precision is that the fraction of retrieved instances that are relevant to the findings.

$$\text{Precision}=\frac{TP}{TP+FP} \tag{25}$$

where,

TP = True Positive (Equivalent with Hits)

FP = False Positive (Equivalent with False Alarm)

7.2 RECALL

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

$$\text{Recall}=\frac{TP}{TP+FN} \tag{26}$$

where,

TP = True Positive (Equivalent with Hits)

FN = False Negative (Equivalent with Miss)

7.3 F-MEASURE

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

$$F_m=\frac{2\times\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{27}$$

8. EXPERIMENTAL RESULTS

The experimental results of the proposed method web data extraction for web document clustering are presented in this section. The projected approach has been enforced within the result shows the output of document classification based on Deep Learning, Naive Bayes and Back Propagation Algorithm. The web URL is that the input and test dataset for this work. And realize content block for this web URL. This work removes the HTML tags, comments, advertisements etc., This is implemented in MATLAB 2013 and therefore the experimentation is performed on a 3.0GHz core i5 computer machine with 4 GB main memory. For experimentation, we have engaged several web pages which contained all the noises corresponding to Navigation bars, Panels and Frames, Page Headers, Footers, Copyright and Privacy Notices, Advertisements and different Uninteresting information. The WebPages are then subjected to method through the proposed deep learning network to web content extraction. The performances of the proposed approach section analyses the exploitation performance measures of precision, recall, and F-measure.

The documents are collected from the online contains both relevant and irrelevant contents as per the necessity 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 group of web documents from the online and manually extracted its blocks with the help of HTML tag. Then they are analysed and therefore the vital blocks are identified. Subsequent to the manual estimation, the proposed deep learning technique is subjected to extract the content of information from web sites.

The content extracted by this technique is compared with content identified manually for evaluating the accuracy of the proposed deep learning based web content extraction.

8.1 ENTROPY ON DEEP LEARING TECHNIQUE

The Deep Learning Technique has selected data set of web documents in the evaluation process. The different blocks are evaluated based on precision, Recall and F-Measures parameter.

Table.1. Deep Learning Algorithm performance analysis

Different Web Pages	Precision (%)	Recall (%)	F_Measure (%)
http://www.international.ucla.edu/korea/	73	75	76
http://www.coronaregional.com/	75	76	79
http://www.pinchin.com/newsletter-list	70	72	75
http://www.medwebplus.com/	84	79	74
http://www.thieme.com/index.php?	82	77	73
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	86	81	82
http://www.eicar.org/	83	82	85

8.2 ENTROPY ON NAIVE BAYSIAN TECHNIQUE

The Naive Bayesian Technique has selected a set of web documents in the evaluation process. The different blocks are evaluated based on precision, Recall and F-Measures parameter.

Table.2. Naive Bayesian performance analysis

Different Web Pages	Precision (%)	Recall (%)	F_Measure (%)
http://www.international.ucla.edu/korea/	78	70	72
http://www.coronaregional.com/	80	74	75
http://www.pinchin.com/newsletter-list	75	67	70
http://www.medwebplus.com/	90	75	69
http://www.thieme.com/index.php?	88	72	67
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	93	77	77
http://www.eicar.org/	95	79	80

8.3 ENTROPY ON BACK PROPAGATION TECHNIQUE

The Back propagation Technique has selected a set of web documents in the evaluation process. The different blocks are evaluated based on precision, Recall and F-Measures parameter.

Table.3. Back Propagation performance analysis

Different Web Pages	Precision (%)	Recall (%)	F_Measure (%)
http://www.international.ucla.edu/korea/	76	73	74
http://www.coronaregional.com/	78	75	78
http://www.pinchin.com/newsletter-list	73	70	73
http://www.medwebplus.com/	89	78	71
http://www.thieme.com/index.php?	85	76	70
http://newsroom.ucla.edu/portal/ucla/the-study-abroad-road-less-traveled-219984.aspx	90	80	79
http://www.eicar.org/	86	81	83

9. COMPARATIVE ANALYSIS

The comparative analysis states the evaluation of the proposed approach with an existing web content extraction method. The three classification algorithm describes document extraction to discover the web content without outliers. In previous method we outlined the Deep Learning and Bayesian for content extraction from web sites.

Table.4. Weighted average values of Deep Learning, Naive Bayes and Back Propagation on evaluation scheme

Techniques	Precision (%)	Recall (%)	F_Measure (%)
Deep Learning	94	74	73
Bayesian	85	65	63
BPNN	89	72	70

10. CONCLUSION

The proposed method is efficient in identifying vital options in content. The proposed approach uses three parameters for the evaluation of the web pages. This paper discusses three web document extraction techniques of Deep Learning algorithms, Naive Bayes Approach and BPNN. The performance metrics are evaluated with comparison of three techniques. The experimental results are recorded in the Table.4. Accordingly the comparative analysis proved that the Deep Learning technique is performed best in all measures of precision, recall and f-measure. This approach yields expected result with accuracy for web content extraction. The weighted average computed for precision, recall and f-measure values are concluded as 94%, 74% and 73% correspondingly. In future, we can incorporate different machine learning technique like hybrid algorithms for capable information extraction from the web sites. And also some other supervised classification techniques will be considered for comparison. Not only will those heuristic algorithms be fused with deep learning algorithm for fast and better improvement.

REFERENCES

[1] Wenyuan Dai, Gui-Rong Xue, Qiang Yang and Yong Yu, “Transferring Naive Bayes Classifiers for Text Classification”, *Proceedings of the 22nd National Conference on Artificial Intelligence*, Vol. 1, pp. 540-545, 2007.

[2] Lawrence McAfee “Document Classification using Deep Belief Nets” CS 224n, Stanford University, 2008.

[3] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, “Automatic Text Summarization using Feature-based Fuzzy Extraction”, *Jurnal Teknologi Maklumat*, Vol. 20, No. 2, pp. 105-115, 2008.

[4] Tak-Lam Wong and Wai Lam, “Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 4, pp. 523-536, 2010.

[5] R.K. Roul and S.K. Sahay, “An Effective Approach for Web Document Classification using the Concept of Association Analysis of Data Mining”, *International Journal of Computer Science and Engineering Technology*, Vol. 3, No. 10, pp. 483-491, 2012.

[6] Jiang Su, Jelber Sayyad Shirab and Stan Matwin “Large Scale Text Classification using Semi-supervised Multinomial Naive Bayes”, *Proceedings of the 28th International Conference on Machine Learning*, pp. 97-104, 2011.

[7] Amit Ganatra, Y. P. Kosta, Gaurang Panchal and Chintan Gajjar, “Initial Classification Through Back Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm”, *International Journal of Computer Science & Information Technology*, Vol. 3, No. 1, pp. 98-116, 2011.

[8] Yan Liu, Sheng-hua Zhong, Wenjie Li, “Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning”, *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 1699-1705, 2012.

[9] Saduf and Mohd Arif Wani, “Comparative Study of Back Propagation Learning Algorithms for Neural Issue Networks”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 12, pp. 1151-1156, 2013.

[10] Citra Ramadhena, Ashraf Osman Ibrahim and Sarina Sulaiman, “Weights Adjustment of Two-Term Back-Propagation Network Using Adaptive and Fixed Learning Methods”, *International Journal of Advances in Soft Computing and its Application*, Vol. 5, No. 2, 2013.

[11] Daniel Soudry, Itay Hubara and Ron Meir, “Expectation Back propagation: Parameter-Free Training of Multilayer Neural Networks with Continuous or Discrete Weights”, *Advances in Neural Information Processing Systems*, pp. 963-971, 2014.