International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

A Web Page Classification Technique with Textual Content Analysis Using NN-PCA for Objectionable Web Page Classification

Deepshikha Patel, Dr. Prashant Kumar Singh

Abstract **– As the Internet has recently been rapidly expanded; we can find information easily and quickly. A lot of useful information exists on the internet, but there is also harmful information involving pornography, adult content, is not appropriate for all users. This is particularly problematic when children are able to access the objectionable material with ease. Pornography web content is one of the biggest harmful resources that pollute the healthy mind of children and teenagers. Several content based analysis approaches had been proposed to avoiding objectionable and other offensive material accessed by the children. Internet users have begun to protect themselves and their wards by using so called web content filters, which allow access to legitimate content and disallow access to objectionable, illegal, pornographic, and other problematic content. This paper, proposes a new content based classification scheme for objectionable web documents filtering. Proposed method uses a neural network with input obtained by Principal Component Analysis (PCA). Each web page is represented by the term weighting scheme. As the number of unique words in the collection set is big, the PCA has been used to select the most relevant features for the classification. These feature vectors are then used as the input to the neural network for classification. We have conducted results by taking four data sets containing different web pages to test the performance of classifier in various scenarios. The experimental evaluation demonstrates that the proposed method provides outstanding classification accuracy for objectionable document classification.**

Keywords **- Classification, Web Document Classification, Feature Selection, Principal Component Analysis, Neural Network.**

I. INTRODUCTION

The World Wide Web is growing ever more rapidly. More and richer information sources and services, such as, news, advertisement, entertainment and consumer information is available on the web. With the rapid penetration of internet will be an important media for future communication and business, perhaps even more important than television. Nowadays internet is the biggest source to access any kind of information that millions of people use as a matter of routine. People start to learn and use internet in their daily activities such as personal communication and entertainment. Companies are actively looking into this new “hot” media, exploring opportunities to increase their company’s productivity and to expand their business. The openness of the web allows any user to access almost any type of information. Given the diversity of views and the ability to post any kind of information on the internet, very often, material that is considered objectionable can be easily accessed on the web. This is particularly problematic when children are able to access these materials with ease. Additionally, for adults, some contents included in abnormal porn sites can do ordinary people mental health harm.

The exponential increase of information on the internet has raised the issue of information security. Objectionable web content is one of the biggest harmful resources that pollute the mind of children and teenagers. This is becoming serious social issue and parents are much concern about the access to this kind of information by children. With the flooding of pornographic material on the web, educators and parents would like to block these offensive materials from their children. The need to protect children and teenagers from objectionable material has led to the development of technologies to facilitate filtering of web content. Now, we see many parents installing objectionable web content filtering software in personal computers.

Apart from these, centralized internet filtering and blocking is also very important to an organizations for a couple of reasons. Companies also want to reduce the amount of work time that its employees spend on non-productive web surfing. The trend starts with the news sites, then entertainment and sport sites and now stock market and currency exchange sites. One important consequences of this phenomenon is that employees start to puzzle whether they should push internet access in the company.

All these result in the great demand for intelligent filtering mechanism that can selectively block information. Organizations and companies further request for centralized solutions so that system administrator, regulation enforcement, and software maintenance can be done in an effective and efficient way.

Currently there is several information blocking or filtering applications designed for use on the web. Examples of such system include CyberPatrol[15], NetNanny[14],Safesurf[16], Honorguard[17], Cyber Snoop[12], CyberSitter[12] etc. Current blocking and filtering mechanisms can be roughly be classified into two approaches. metadata based method and content based method. Metadata based method is based on URL/IP address blocking. Most of the filtering solution blocks connection to URLs [35], [36], which are in URL database. In this a requested URL address will be blocked if a match is found in block list. This is known to be more effective and efficient. But, as the URL/IP of web sites are always changing; New sites are kept uploading onto the internet daily; the sites might also be moved regularly. However keeping the list up-to-date and generation and maintenance of the URL database is very difficult, slow

and time consuming task. Not only do many web sites appear every day while other disappears, but also site content is updated frequently. Thus, manual classification and filtering systems are largely impractical. As a result the highly dynamic character of the web calls for new techniques designed to generate and maintain the URL database by automatically collecting and classifying web contents. Currently content based approaches are new trend for web filtering and blocking. Content analysis based approaches involve deeper understanding of the semantics of text and other media items especially multimedia contents, by using linguistics analysis, machine learning, and image processing components.

II. Related Work

Classification of objectionable documents is an important research area these days. For example Reference [41] uses multilevel classification of objectionable content using SVM. WebGuard, a hybrid web filtering system [11] used classical data mining techniques on prohibited keywords appear on a test web page to make classification decision in its textual analysis component. Reference [10] proposed early decision heuristics for objectionable content classification using an inverse chi-square classification. Reference [6], classified the objectionable texts into four rates according to their harmfulness and proposed the hierarchical text rating system for objectionable documents. Reference [1], uses a modified entropy term weighting scheme as feature selection for illicit web page classification. In this paper [1], they try to solve similarly content issue in pornography and medical consultant web page. Reference [3], investigate how far the multimedia content analysis should go for Internet filtering and blocking. How to efficiently utilize the bandwidth is a great challenge to network administrators and Internet users. In order to improve the efficiency of Internet bandwidth utilization, [38] proposes a web-based P2P content classification approach to objectionable content filtering.

With the flooding of inappropriate information, on the internet, how to keep people especially children and teenagers away from that offensive information is becoming one of the most important research areas these days [1]-[3],[5]-[9]. On the other hand, end users themselves have expressed their need for filtering technologies[13], for example parents request better technical tools for protecting their children in the internet [14]-[17], in particular for filtering pornography. To filter out objectionable content on the internet there is a need for a tool which automatically collects these offensive materials from the web. For this purpose Reference [37], proposed a specialized web robot to automatically collect objectionable web content for use in an objectionable web content classification system.

III. Problem Statement

Classifying and filtering the wide spreading objectionable web content has attracted intensive attention on protecting children and anyone else from access to them[9]. Controlling access to the objectionable web content typically employs different approaches including meta-data based approach and content based approach.

The meta-data based approach depends on the results of URL and IP addresses blocking. Because the internet is very dynamic and the URL/IP addresses of web sites are always changing, this approach has the problem of having to periodically update blocking list. The highly dynamic character of the web calls for new techniques designed to classify and filter web sites and URL automatically. To solve such problems several studies of content based filtering have been conducted [35],[36].

There are various commercial web filtering products available in the market to overcome the issue of illicit web content. Some popular products are CyberPatrol, NetNanny, Websense, and etc. [12]. Unfortunately their filtering effectiveness is limited by the use of several traditional techniques such as URL (Uniform Resource Locator) blocking, PICS (Platform for Internet Content Selection) and keyword matching. These techniques are advantages from easy to implement and require only short processing time. As a result, these techniques fail to identify the web pages sufficiently when the web content changes from time to time.

Currently content based analysis approach becomes a new trend for web filtering research [8],[39]. Content analysis is an approach that involves a deeper understanding the semantics of text and other media items (especially pictures), by using linguistic analysis, machine learning, and image processing components. But these methods also have some shortcomings. In text stream analysis one of the main problems is finding an effective method to classify documents fast and correctly. In such applications the dimensionality of term space may be problematic. The classifier performance degrades with the higher dimensions.

Another problem with content based filtering is mis-blocking. In content filtering approach, keyword matching [20] is often used. Many desirable web sites are blocked because some predefined keywords appear in their web pages, though in different meaning or context.

Most of the approaches are weak against classifying high similarity web content such as pornography and genecology web pages. Similarity content issue refers to web pages that share the similar terminology for several topics though in different meaning or context. Due to the high similarity of the terms features, most of the classifier misclassify the objectionable web pages as normal/non objectionable web pages or vice versa.

IV. Proposed Solution

Currently no perfect web classification design has been found, as each classification design is highly dependable on the content of the web site. Web site classification is an ongoing process prone to error. Each time a new document of content is published on the web site, it needs to classify. If the document is classified wrongly, then it undermines the entire classification design. Therefore, various

approaches have been applied to automated web page classification in order to improve its performance.

Proposed method focuses on two major problems with content based filtering, discussed below and try to overcome them.

Problem1: Dimensionality of term space

When the dimensions of the original feature vector are large, this may be problematic. High dimensionality of feature space may slow down the classifier and the classifier performance may degrade.

Solution: Using PCA as feature reduction

Above problem is overcome in proposed method using Principal Component Analysis (PCA). Proposed web page classification method uses PCA to reduce the dimensions of original feature vector, thus improve the classification performance.

Problem2: Mis-blocking of websites

Most of the approaches are weak against classify the high similarity web content such as pages related to pornography and medical consultation. Because of the similarity issue, many authorized web sites are blocked i.e. mis-blocking.

Solution: Considering length factor in Entropy term weighting scheme

Pornography web pages normally contain few numbers of words and short in document length. In fact length of the document should be one of the consideration factor in Entropy if implement for illicit web page classification especially handling similarity issues.

V. Objectionable Web Page Classification Model

There is always a challenge to classify objectionable and non-objectionable web pages. In order to meet the goal of classifying objectionable web pages, this work proposes the Objectionable Web Page Classification Method (OWPCM) by implementing new term weighting scheme and PCA as feature reduction process. Basically the design of OWPCM consist of five main parts which are *web documents collection, pre-processing, feature selection & Representation, re-parameterization,* and *classification* parts depicted in Fig 1.

A. Web Document Collection

Web document collection is the process to retrieve web pages from the different web sites with the help of web browser. Those retrieved web pages are then stored in the local database for further processing.

B. Pre-processing

All collected web documents in document collection process are pre-processed in this part of classification process. All web documents are HTML documents and contain HTML tags, so they are HTML parsed so that only texts are extracted excluding those HTML code [21],[22],[24],[25]. Afterward stopping and stemming are performed on those documents. Stopping is a process of removing most frequent words that exist in a web document by using a stop words dictionary. However stemming reduces the occurrence of term frequency,

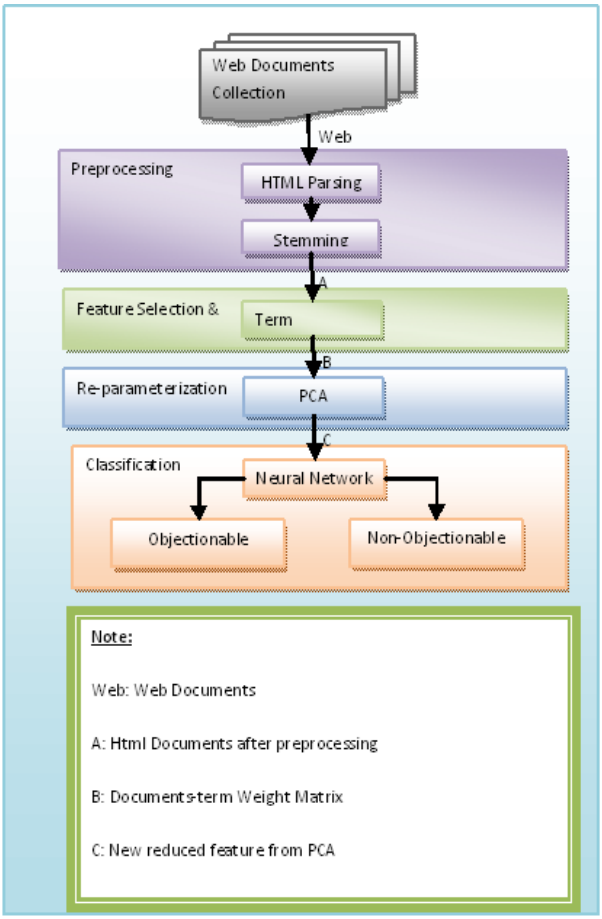

Fig.1. Objectionable Web Page Classification Model

which has similar meaning in the same document [32]. For this reason, a number of so-called *stemming Algorithms*, or *stemmers*, have been developed, which attempt to reduce a word to its *stem* or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be *conflated* to a single representative form – it also reduces the *dictionary size*, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

In the proposed model, Porter Stemming Algorithm [23] is used. The Porter Stemmer is a conflation Stemmer developed by Martin Porter at the University of Cambridge in 1980. The Stemmer is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes.

C. Feature Selection and Representation

Feature selection is a process which selects features from original set of features [26]-[31].In OWPCM, modified entropy term weighting scheme is used as feature weighting.

1) Entropy Term Weighting Scheme

Entropy method is based on a probabilistic analysis of the texts. Entropy believes that significance of term is proportional to the frequency of a term in most documents. Term can be assigned weights according to local term

weighting and global term weighting. In objectionable web pages classification scenario, sex consultation, Genecology web pages sometimes may contain similar offensive terms and may be classified in the same category of documents. Entropy will not work well in this type of classification. In a sense that sexology and genecology web pages always having larger length of document, contrast to illicit web pages which normally contain few number of words and short in document length. So we can consider the length of document as an important factor in Entropy when implement for objectionable content classification specially when we need to classify several web pages including sex education, gynaecology and other illicit web pages.

2) Modified Entropy Term Weighting Scheme

The Text categorization problems normally involve an extremely high dimensional feature space. A standard procedure to reduce features dimensionality is feature selection. We will use modified entropy scheme to extract features from text documents. The term with highest term weights would select as best features. Later, these selected features would form as feature vector. The modified term weighting scheme [1] is as follows.

$$G_i = \left(\frac{\log_{10} DF_i}{\log_{10} n} + 1 \right) \quad (1)$$

$$L_{ij} = \begin{cases} K_{ij} \ (TF_{ij} > 0) \\ 0 \ \ (TF_{ij} = 0) \end{cases} \quad (2)$$

$$K_{ij} = \left(\frac{\log_{10}(TF_{ij})}{\log_{10}(lenDoc_j)} + 1 \right) * \left(\frac{\log_{10}(TF_{ij})}{\log_{10} T(i)} + 1 \right),$$

$$T(i) = \sum_{j=1}^{n} TF_{ij} \quad (3)$$

$$W_{ij} = L_{ij} \times G_i \quad (4)$$

Where,

DF_i = no. of documents that contain *i*th term in a collection.

$lenDoc_j$ =no. of total words that exists in *jth document* (but not the total number of unique word that exist in *j*th document).

This modified term weighting scheme contains three main factors including term frequency, collection frequency and document length factor. More detailed discussion about modified entropy term weighting scheme can be found in the work of Lee et al. [1].

The modified term weighting scheme is originally improved from Entropy where to suit the purpose for objectionable web pages classification. The improved term weighting is considering three main factors which are term frequency factor, collection frequency factor, and document length factor. Improved Entropy term weighting scheme considers the document length term and frequency factors by making assumption that the length of pornography web pages text is mostly shorter than sexology web pages. The length (not unique term) of a web page text should be one of an important factor that effecting to illicit web page classification performance. Improved Entropy term weighting scheme try to identify the objectionable and offensive terms by strengthening the term weight if it is a short document, weaken the term weight if it is a long document. Hence, the frequent occurrence of a term is divided by document length, where $\left(\frac{\log_{10}(TF_{ij})}{\log_{10}(lenDoc_j)} + 1 \right)$ will penalize the term weight if the document length is longer. On the other hand, it also considers the term frequency factor by computing the distribution rate of a term within the collection via $\left(\frac{\log_{10}(TF_{ij})}{\log_{10} T(i)} + 1 \right)$ and the condition of ($TF_{ij} > 0$) must be fulfilled. The higher occurrence of the term (TF_{ij}), the bigger denomination it is. Hence the local term weight (L_{ij}) could be obtained by considering both factors as illustrated at Eq.(3) and Eq.(4). In fact, the improved entropy term weighting argue that length document should be consider as a factor during calculating a term weight which previously exclude in Entropy term weighting scheme.

D. Re-parameterization Using Principal Component Analysis

Using PCA, the dimension reduction process will reduce the original data vector into small number of relevant features [4],[33],[34],[42]. PCA allows us to compute a linear transformation that maps data from a high dimensional space to a lower dimensional space. The best low-dimensional space can be determined by the "best" eigenvectors of the covariance matrix of x (i.e., the eigenvectors corresponding to the "largest" eigenvalues also called "principal components"). There are five basic steps needed to perform a Principal Component Analysis are shown below:

Step 1: Get some data

This is the data on which, PCA is to be applied i.e. document-term weight matrix.

Let M to be the matrix of document terms weights as follows.

$$M = \begin{pmatrix} x_{11} x_{12} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

Where,

x_{jk} = term weight.

j= variable, i=1,2,.....,n

k= variable, j=1,2,.....,m

Step 2: Subtract the mean

For PCA to work properly, we have to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. This produces a data set whose mean is zero. The mean of m variables in data matrix M will be calculated by

$$\overline{x_k} = \frac{1}{n} \sum_{j=1}^{n} x_{jk} \quad (5)$$

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Step 3: Calculate the covariance matrix

After that the covariance matrix C= $\{c_{jk}\}$ is calculated. The covariance of matrix c_{ik} is given by:

$$c_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x_i}\right)\left(x_{jk} - \bar{x_j}\right) \quad (6)$$

Where, i=1,2,...m

Step 4: Calculate the eigenvalues and eigenvectors of covariance matrix

Since the covariance matrix is square, we can calculate the eigenvalues and eignvectors for this matrix. These are rather important, as they tell us useful information about our data. An eigenvalue λ and eigenvector **e** can be found by C**e**= λ**e** where C is covariance matrix. If C is an m x m matrix of full rank, m eigenvalues and all corresponding eigenvectors can be found by using

$$(C - \lambda_i I)e_i = 0 \quad (7)$$

Step 5: Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into it. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalues, highest to lowest. This gives you the components in order of significance. In fact, it turns out that the eigenvectors with the highest eigenvalue is the *principal component* of the data set. Now feature vector is constructed by taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the column.

Step 6: Deriving the new data set

This is the final step in PCA, and is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

E. Neural Network as Classifier

For many years, there was no theoretically sound algorithm for training multilayer artificial neural network [40]. Since single layer network proved severely limited in what they could represent, the entire field went into virtual eclipse. The resurgence of interest in artificial neural network began after the invention of backpropagation algorithm. In OWPCM, most popular artificial neural network (ANN) architecture, the Multilayer Feedforward (MLFF) network with backpropagation (BP) learning is used. This type of network is sometimes called multilayer perceptron because of its similarity to perceptron network with more than one layer.

Back propagation learning

Backpropagation neural networks [40] employ one of the most popular neural network learning algorithms, the Backpropagation (BP) algorithm.

The basic algorithm loop structure is given as:

Initialize the weights
Repeat
    For each training pattern
    Train on that pattern
    End
Until the error is acceptably low

VI. EXPERIMENTS

A. Data

The first task is to build a suitable test corpus of web pages. It is the very first step of any classification model. We have collected more than 1000 web pages of the categories objectionable and non-objectionable with the help of search engines for the further processing. We only have chosen around 712 documents to form the dataset. These collected documents contain 550 objectionable and 162 non objectionable documents. Out of those selected documents, randomly 649 documents are selected for training sets and 63 documents are selected for test sets. However the datasets that are being used for non-objectionable category contain medical, computer, and sports web pages. We prepared four Datasets to test the efficiency of OWPCM. All documents of training sets for all four datasets are manually categorized into "objectionable" and "non-objectionable" classes, because the proposed model will classify the web documents into two categories namely objectionable and non-objectionable and test set contains combined documents of both the categories. Training documents and test documents are stored in local database. Table: 1 summarizes the dataset. The dataset is divided into four type of dataset which are illustrated in Table:2. The purpose to do so is to examine the identification performance of OWPCM from four different situations.

Table 1: Summary of data set

No.	Category	Label	Total Web Pages
1	Objectionable	Obj	550
2	Computer	Comp	34
3	Medical	Med	48
4	Sports	Spt	80
Total			712

Table 2: Train sets and test sets

Dataset	Categories Combination	No. of Pages in Train set		No. of Pages in Test set	
		Category	Pages	Category	Pages
Data set 1	Obj + Comp	Obj	121	Obj	10
		Comp	15	Comp	04
Data set 2	Obj + Med	Obj	139	Obj	11
		Med	22	Med	04
Data set 3	Obj + Spt	Obj	185	Obj	12
		Spt	55	Spt	03
Data set 4	Obj + Comp + Med + Spt	Obj	67	Obj	05
		Comp	11	Comp	04
		Med	17	Med	05
		Spt	17	Spt	05
Total		649		63	

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Table 3: Features selected

Dataset	No. of Pre-processed Web Pages	No. of Features Selected
Data set 1	136	4079
Data set 2	161	4626
Data set 3	240	6208
Data set 4	112	3715

B. Experiment Environment

In the proposed method, feed forward back propagation neural network with n inputs and 2 outputs is used for the classification purpose, where n is the total number of output features from the PCA. This module involves two steps: neural network learning and classification of new web document.

First step is to train the network with training set documents. Then output of the PCA is feed to the neural network as input. After this process, network is now ready to classify documents into the recognized class based on the learning. In order to do the classification, we implement feedforward backprogation neural network. We have used four kind of data sets as shown in Table: 2 and parameters specifying BP-ANN are summarized in Table: 4. We found that the network will achieve an optimum result if given parameters' values are as shown in Table: 4. Figure: 2 represents the architecture of BP-ANN used in OWPCM.

Table 4: BP-ANN parameters used in OWPCM

Parameters	Values
No. of Input node (I)	Variable(n)
No. of output node (O)	02
No. of Hidden node (H)	n/2
Learning Rate (l)	0.5
Momentum (m)	01
No. of Interaction (i)	100

Where,

I= No. of output from PCA

H= I/2= (No. of output from PCA)/2

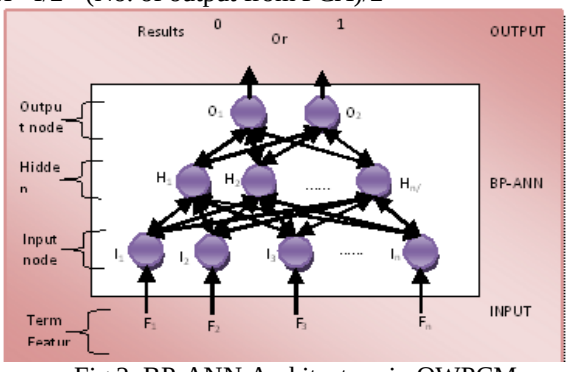

Fig.2. BP-ANN Architecture in OWPCM

In this paper, we select output weights from PCA as input to Neural Network. Suppose total no. of output from PCA is 'n', i.e. number of input to BP-ANN is 'n'. In this implementation we designed two output neurons. Result from neurons is {0,1} or {1,0}. {0,1} represents the "Non-objectionable" web page and {1,0} represents the "Objectionable " web document. Thus the number of input neurons would be I=n, number of hidden neurons is H=n/2, and number of output neurons is O=2.

A. Performance Evaluation

After the classification predictions come out, a series of performance measures can be employed. Namely, recall, precision, F1, accuracy and error [10], [18], [19], [32].

For each category or class, these measures can be defined as follows:

Table 5: Contingency table for category c_i

	Expert Yes	Expert No
Prediction Yes	a_i (Assigned correctly)	b_i (Assigned wrongly)
Prediction No	c_i (Rejected wrongly)	d_i (Rejected correctly)

$$recall = \frac{a_i}{a_i + c_i} \quad (8)$$

$$precision = \frac{a_i}{a_i + b_i} \quad (9)$$

$$accuracy = \frac{a_i + d_i}{a_i + b_i + c_i + d_i} \quad (10)$$

$$error = \frac{b_i + c_i}{a_i + b_i + c_i + d_i} \quad (11)$$

$$F1 = \frac{2 * recall * precision}{recall + precision} \quad (12)$$

The F1 score is the (equally weighted) harmonic average of recall and precision. In addition, we define the false positive rate and false negative rate as

$$\text{false positive rate} = \frac{b_i}{d_i + b_i} \quad (13)$$

$$\text{false negative rate} = \frac{c_i}{a_i + c_i} \quad (14)$$

Here the letter 'a_i', 'b_i', 'c_i' and 'd_i' are from Table 5. *i,* is the total number of categories. a_i,is the number of documents correctly assigned to the category C_i; b_i is the number of documents incorrectly assigned to the category C_i; c_i is the number of documents incorrectly rejected from category C_i; and d_i is the number of documents correctly rejected from category C_i. Moreover, $a_i + c_i$ represents the total number of documents that belong to the category C_i; $a_i + b_i$ indicates the total number of documents assigned to the category C_i; $b_i + d_i$ represents the number of documents that should not be in the category C_i; and $a_i + b_i + c_i + d_i$ is the total number of documents evaluated for the category C_i. We can make contingency table for each category.

VII. RESULTS

We have conducted experiments on four data sets to test the classification performance. We have compared the

results from all data set in aspect of average of precision, recall, F1, accuracy rate, Error rate, false positive rate and false negative rate. In all cases our model has achieved outstanding results. Table: 6 shows the average of all performance measures considered for both categories of data set 1, data set 2, data set 3, and data set 4 respectively. For data set 2 and data set 3 our model has performed 100%. Figure 3 and Figure 4 shows that proposed model has achieve better performance aspect of all performance measures. It is also imply that proposed classification model is excellent for classifying objectionable web pages under all four cases which are Data set 1(Computer related web pages combined with objectionable web pages), Data set 2 (Medical consultation web pages with objectionable category), Data set 3 (objectionable web pages with sports web pages), and Data set 4 (mixed web pages from all four categories).

Table 6: average of performance measures for all data sets

Category	Data set 1	Data set 2	Data set 3	Data set 4
Precision	95.45%	100%	100%	91.67%
Recall	87.5%	100%	100%	96.43%
F1	90.47%	100%	100%	93.6%
Accuracy	92.86%	100%	100%	93.6%
Error	7%	0%	0%	5%
False positive rate	125%	0%	0%	3.5%
False negative rate	125%	0%	0%	3.5%

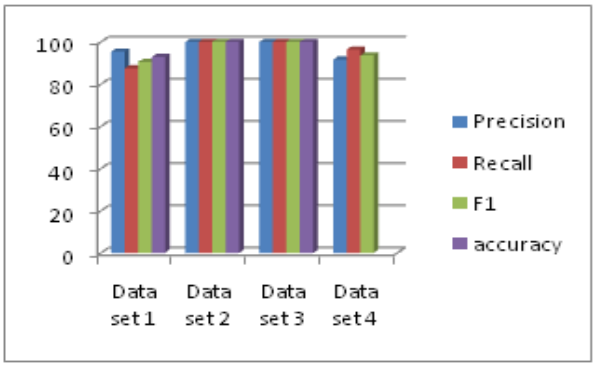

Fig.1. Comparison of average of Precision, Recall, F1, and accuracy

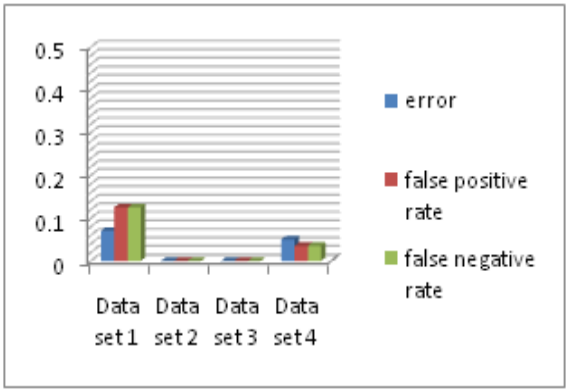

Fig.2. Comparison of average of error, false positive rate, and false negative rate

VIII. Conclusion

In this paper, we designed and implemented a web page filtering system for objectionable web documents. We proposed OWPCM for classifying objectionable and non-objectionable web documents. This system was experimented using the NN machine learning algorithm, entropy term weighting indexing algorithm, and PCA feature reduction algorithm. We have tested the performance of our classification system by taking four different types of data sets. Our experimental results have achieved great accuracy for classifying web documents. It is really a challenging task to categorize web pages having similar contents. Our proposed model is good against classifying these web pages. Using this model in other application are would be further in future.

References

[1] Z. S. Lee, M. A. Maarof,A. Selamat, S.M. Shamsuddin, "Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classofication,"*IEEE 8th Int. Conference on Intelligent System Design and Applications*, 2008, pp.145-150.

[2] L.H. Lee, C.J. Luh and C.J. Yang,"A study on Early Decision Making in Objectionable Web Content Classification,"*IEEE*, 2008, pp. 35-39.

[3] C. Ding, C.H. Chi, J. Deng, and C.L. Dong,"Centralized Content-Based Web Filtering and Blocking: How Far Can It Go?" *IEEE SMC'99Conference Proceedings*, 1999, pp.115-119.

[4] Ali Selamat, Hidekazu Yanagimoto and Sigeru Omatu," Web News Classification Using Neural Networks Based on PCA,"*SICE02-0163.*

[5] Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank,"Recognition of Pornographic Web Pages by Classifying Texts and Images," *IEEE Transactions On Pattern Analysis And Machine Intellegence, Vol. 29, NO. 6*, June 2007, pp. 1019-1034.

[6] C.Y.Jeong, S.W. Han and T.Y.Nam,"A Hierarchical Text Rating System for Objectionable Documents,",*Int. Journal of Onformation Processing System Vol.1,No.1*, 2005, pp. 22-26.

[7] K. V. Chandrinos, Ion Androutsopoulos, G.Paliouras, and C. D. Spyropoulos," Automatic Web Rating:Filtering Obscene Content on the Web,", *ECDL 2000, LNCS 1923*, pp. 403–406, 2000.

[8] Z.S. Lee, M.A. Maarof, A. Selamat. S. M. Shamsuddin.:"Pornography Web Pages Classification with Textual Content Analysis Using Entropy Term Weighting Scheme for Small Class Dataset", *Proceeding of 3rd Postgraduate Annual Research Seminar 2007 (PARS'07), Universiti Teknologi Malaysia*, 2nd-5th July 2007.

[9] V. Jacob, R. Krishnan, Y. Ryu, R, Chandrasekaran and S. Hong, "Filtering Objectionable Internet Content," *in Proc. of the 20th International Conference on Information Systems* , 1999, pp. 274-278.

[10] Jiawei Han, Micheline Kamber," Data Mining : Concepts & Techniques ," *Morgan Kaufmann publications second Edition*, 2006.

[11] G.Y. SU, J.H. LI, Y.H. MA and S.H. LI, "Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model," *Journal of Zhejiang University Science, 2004 Vol. 5 No. 9*,2004,pp.1106-1113.

[12] "Internet Filter Review", available at http://www.internet-filter-review.toptenreviews.com/, May 2007.

[13] G. McGovern, 2001. Web Classification is Essential, available at http://www.gerrymcgovern.com/nt/2001/nt_2001_11_26_classify.htm

[14] Trove Investment Corporation, "Net Nanny: the best way to protect your children and free speech on the Internet", 1995. http://giant.mindlink.net/netnanny/home. html.

[15] Microsystems Software, "CyberPatrol", August 1999. http://www.microsys.com/cyber/default. htm
[16] SafeSurf. http://www.safesurf.com/.
[17] Honorguard. Christian Filtered Internet Service on the Internet. http://www.honorguard.net/.
[18] F.S ebastiani.: "Machine Learning In Automated Text Categorization", *ACM Computing Surveys, vol 34, no.1*, pp. 1-47, 2002.
[19] Fabrizio Sebastiani ,"A Tutorial on Automated Text Categorisation," *Istituto di Elaborazione dell'Informazione Consiglio Nazionale delle Ricerche Via S. Maria*, 46 - 56126 Pisa (Italy).
[20] H. Mase and H. Tsuji.:"Experiment on Automatic Web Page Categorization for Information Retrieval System,"*Journal of Info. Processing Society of Japan, Vol. 42,No.2*, 2001,pp. 334-347.
[21] D. Freitag, "Information Extraction from HTML: Application of a general machine learning approach,", *In Proc. Of 15th Nat. Conf. AI* ,1998.
[22] F. R. Rahman, H. Alam and R. Hartono,"Content Extraction from HTML Documents'",pp. 7-10.
[23] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3,pp. 130–137, 1980.
[24] S.Soderland.: "Learning to extract text-based information from the world wide web", *In Proceedings of 3rd International Conference in Knowledge Discovery and Data Mining (KDD-97)* , 1997, pp. 251-254.
[25] Jan and Radim,"The Influence of preprocessing parameterson text categorization World Academy of Science,", *Engineering and Technology*, 2007,pp. 54-57.
[26] Yang, Yiming, " A Comparative Study on Feature Selection in Text Categorization," *In Proceeding of the Fourteenth International Conference on Machine Learning(ICML'97)*, 1997,pp.412-420.
[27] X. Qi and B. D. Davison," Web page classification: Features and algorithms,", *ACM Computing Surveys*, 41, 2009.
[28] D.D.Lewis, "Feature selection and feature extraction for text categorization,", *Proceedings of Speech and Natural Language Workshop*, 1992.
[29] Yang, Y., Liu, X.,"A Re-examination of Text Categorization Methods,", *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999,pp.42-49.
[30] Tao Liu Shengping, Liu Zheng Chen," An Evaluation on Feature Selection for Text Clustering,", *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.
[31] M. Rogati and Y. Yang, "High Performance Feature Selection for Text Classification," *In Proceeding of CIKM'02,Melean, Virginia, USA*, 2002, pp.659-661.
[32] Z. Markov, Daniel T. Larose,"Data Mining The Web,", Wiley-Interscience Publication, 2007.
[33] Lindsay I Smith,"A Tutorial on Principal component Analysis,", February 26, 2002
[34] Lee Zhi Sam, Mohd Aizaini Maarof, Ali Selamat," Automated Web Pages Classification with Integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as Feature Reduction,", *Proceedings of International Conference on Man-Machine Systems 2006*, September 15-16, 2006, Langkawi, Malaysia.
[35] L. K. Shih and D. Karger,"Using URLs and table layout for web classification tasks,", *In Proc. of WWW '04*, 2004.
[36] M. Kan and H. O. N. Thi. ,"Fast webpage classification using url features,", *In International conference on Information and knowledge management (CIKM)* , 2005, pp.325-326.
[37] SuGil Choi, Seung Wan Han, Chi-Yoon Jeong, and Taek Yong Nam ,"Specialized Web Robot for Objectionable Web Content Classification,", *World Academy of Science, Engineering and Technology 7*, 2005,pp.18-21.
[38] Jenq-Haur Wang1, Hung-Chi Chang1, Ming-Jer Lee2, Yu-Mei Shaw3 "Classifying Peer-to-Peer File Transfers for Objectionable Content Filtering Using a Web-based Approach,".
[39] Michael G. Noll, Christoph Meinel," Web Page Classification: An Exploratory Study of the Usage of Internet Content Rating Systems,", *HACK 2005* Luxembourg City, Luxembourg.
[40] S.Rajshekharan, G. A. Vijayalaxmi, "Neural Network, Fuzzy Logic, and Genetic algorithms Synthesis & Applications," PHI publication,2007.
[41] Y. Kim, T. Nam,and D. Won,"2-wat Text Classification for Harmful Web Documents,".
[42] R. A. Calvo, M. Partridge, and M. A. Jabri, "A Comparative Study of Principal Component Analysis Techniques", *presented at In Proc. Ninth Australian Conf. on Neural Networks, Brisbane*,1998.

AUTHOR'S PROFILE

Deepshikha Patel

born in Hoshangabad (M.P.) in India, on April 23, 1984. She completed his Engineering in Information Technology from Jabalpur Engineering College, Jabalpur, Madhya Pradesh in 2006 and M. Tech in Computer Science & Engineering from Technocrats Institute of Technology, Bhopal in 2010. Presently, she is pursuing PhD in Computer Science & Engineering from Dr. K. N. Modi University, Newai, Rajasthan and working on customized search engines for kids for her thesis work. She worked as an Assistant Professor in Information Technology Department in Technocrats Institute of Technology, Bhopal, Madhya Pradesh. She is member of various technical societies like IACSIT, IAENG, CSTA, IAOE, UACEE, ISOC, SSRGJJ, AIRCC, SCIEI and British Science Association etc.