WEB PAGE CLASSIFICATION BASED ON NOVEL BLACK WIDOW META-HEURISTIC OPTIMIZATION WITH DEEP LEARNING TECHNIQUE

Manikandan N1, Ruby D2, Gayathri A3, Gopalakrishnan T4

1Associate Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India – 632014, Email: mkyadhav@yahoo.com
2Assistant Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India – 632014, Email: ruby.d@vit.ac.in
3Assistant Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India – 632014, Email: gayathri.a@vit.ac.in
4Assistant Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India – 632014, Email: gopalakrishnan.ct@gmail.com

ABSTRACT

Classification of websites is an application for information recovery that offers useful information, and can form the basis for numerous application fields. The classification of websites offers valuable knowledge for effective use of the Internet, spam filtering and many other fields of use. Rapidly finding important findings from millions of websites is a significant issue that search engines need to address. There are a large number of Web pages that have features like HTML tags, URLs, hyperlinks or text content to take into account during an automated grading process. The current listing methods have several disadvantages such as interfacing with regular website operations and crawling loads of inutile data. The purpose of this study is to lessen the sum of features to progress website classification runtime and accuracy. In this survey, we have selected the best features with a Novel Black Widow Meta-heuristic Optimization (NBW-MHO) algorithm and have then proposed a deep learning classification model which can acquire the best features on each web page and helps in listing search engines. Our proposed algorithm offers optimum classification of websites, which have several web pages, taking into account basic information such as link text, side information and header text. In our experiments, we used the WebKB data set and showed that the use of the NBW-MHO for selecting features enhances both accuracy and classification efficiency. We also found that with detail to well-known knowledge gain and chi square selection approaches, the NBW-MHO-based algorithm can better select features.

Index Terms: Feature Selection; Optimal Classification; Link Text; Spam filtering; Web page classification.

I. INTRODUCTION

With the unbelievable development of the World Wide Web (known as the Web), the website with satisfactory information is difficult to find and unnecessary and damaging content is filtered out. In the last decade, however, there has been a widespread growth of Web pages with harmful and offensive content such as fraud, photography, crime, radicalism, cyber attacks, porn, etc. The large number of web pages with various subjects, on the other hand, has prevented the acquisition of knowledge and extraction models from providing optimal topical findings. In addition to the expertise of machinery learning models for classification of text and image, the explosive efficiency and memory space growth in computer machinery have paved the way for automated solution of complex semantical problems that until a decade ago seemed unrealistically to any computer [1,2]. Semantic web pages based on their content are one of these issues. Website classification involves the assignment to more predefined categories of a website which plays an important part in focused crawling, supporting Web directory creation, topical web connection analysis, background publicity and analyzing the topical structure of the website.
In many recovery and management activities, text classification plays a critical role, including the collection of information, the extraction of information, filtering of documents and the development of hierarchical directs [3]. The web classification or classification of the webpage is referred to if the texts are classified on the web pages. Web pages, however, are diverse from text and comprise a large number of additional material, such as URLs, links, HTML tags that are not backed up by text documents. This web-page property differs from the classification of standard text [4]. Classification on the Internet is used for the selection of the basic topic of web links, the study of the current web structure, web directory creation and focused crawler [5]. In the past, certain web directories such as Yahoo! [6] and the Open Directory Project [7] had been manufactured manually, and web documents were manually assigned class labels. The manual classification, however, takes time and requires a lot of human work, so it is unsalable with regard to the rapid growth of the network. Automated web page classification systems were therefore highly needed [8].

Web classification algorithms are used in many different domains, including spam detection, web search, organization of documentation and cyber security [9]. In this paper, we intend to search the web and optimize the search engine to provide users with rapid and reliable search results. The web classification method is as follows, first, to fix the sum of classes and the features of each class; second, to determine the likelihood of any class for each document, given the collection of training documents; and finally, to determine the class that has the highest probability. There are several web classification learning algorithms, such as Naive Bayes Support Vector Machine (SVM), PSO and k-Nearest Neighbours (kNN) [10]. Most of the model machines, however, do not achieve the desired accuracy since several features exist in a single web page. A good function selection algorithm and the ML model are therefore combined to improve the precise classification process.

The selection of features (FS) is a key step before the classification process. It includes the detection, for the classification method, of a subset of relevant characteristics (significant word). Due to the broad document size [11], this is very required prior to the classification process. Its key benefits include promoting information comprehension, reducing training periods and solving the issue of dimensionality. Furthermore, complications with the classifier and requirements for handling (e.g. storage and office space) are being reduced [12]. The objective of an algorithm for feature selection is to eliminate most irrelevant features of the Web page in order to minimize the input size fed to the system model. As the size of the input is small, a ML model makes it easier to learn how different features correlate and performs more precisely.

The rest of the paper is arranged: Section 2 discusses limiting current web page classification techniques. Section 3 provides an explanation for the projected approach, which describes the validation of the proposed method with regard to current techniques in section 4. Lastly, Section 5 comprises the conclusion of the research work.

II. LITERATURE REVIEW

In this subparagraph we provide a brief overview of the recent research on link-based algorithms for Web sorting. The main purpose is to use the clustering algorithm based on distance calculation steps in order to minimize the average distance between websites of the same class and to maximize the average distance between different websites of different classes. [13] describes the use of a web page tokenization for extracting features; then, a web page class is allocated based on the distance between features calculated. Some researchers use ML algorithms for categorizing web pages which draw large sum of web links, for example, to support vector machines [14].

Few pieces such as [15] quantify the value of fitness for the document classification for each iteration. In each subsequent iteration, the fitness function becomes more precise and the iteration ends when the fitness function is no more increased. There are several works in the swarm optimization field, which are an algorithm based on nature where each bird member of a group seeks food in many places and converges when it finds food. There are several different works. Using the same algorithm, web documents are classified by seeking links through web page documents, and ultimately, all partial solutions are combined into a super optimal solution [16, 17].

In this field, the majority of work is focused on the word bag model [18]. The document's links are then transformed to Vectors, while the most frequent terms on the webpages are extracted using few algorithms, including the term frequency-inverse document frequency (TF-IDF) [19]. TF-IDF is only effective if each word in the document matrix is separate from the other, if two or more words are synonymous each other, then the exact relationship is difficult to represent; research [20] suggests a technique which calculate the correlation

www.turkijphysiotherrehabil.org
between each pair of words and the other. The frequent words are often used as features in classification of web pages [21-23] proposes an algorithm based on kNN which uses a TF-IDF metric for Lao text classification, and is widely used in the natural language processing of the study (NLP). Seven attributes of the input size used in your analysis. The main factor analysis is the feature selection approach they use. Only 3 characteristics are deliberated according to the PCA method of selection and the accuracy of their projected work is 71.4%.

Optimizes the multi-class SVM. [24] SVM is fine if there are only two target classes, i.e. a binary classification, but the cost of computational training is increased when there are more than 3 classes. They proposed in their research a hierarchical grading model in which the SVM is optimized for a multiclass grading process. Their work was contrasted with decision tree classifiers and kNN classifiers and their performance improved. In the SVM classification several optimization [25, 26, 27] were performed for better multi-class classification performance.

In the classification of websites, AntMiner[28] is the first research using the ACO. Inspired by AntMiner[28], Holden and Freitas [29] used the paradigms of an ant's colony to find a set of rules which classify webpages into various classes. They have no previous hypotheses that terms can be used as possible discriminators on the web pages to be graded. They use a stemming technology to minimize data rareness, which considers different grammatical types of a root word equal to aid, assist and help. The best result from AntMiner is a classification accuracy of 81.0 percent, which is achieved when WordNet generalization with title functions is used.

The ACO-based algorithm for text classification was proposed by Aghdam et al. [30]. The functionality selected by an ant is evaluated by the classifier output and the subset duration of the feature. In the experiments, classification efficiency is considered to be larger than subset length, so that 80% and 20% of weights are assigned to classify performance and the subset length respectively. A simple k-NN classifier is used in the experiments to test classification performances. In comparison to the performance of a genetic algorithm, knowledge gain and chi square analysis of feature selection in the Reuters 21578 dataset, the performance of the proposed algorithm is measured. Your experimental assessment demonstrated the superiority of the genetic algorithms, data gain and methodological analysis in the ACO-based function segment. They only analyzed a selection of terms of the extraction method and noted 89.08% of the Reuters 21578 data set's micro-average F-measurement values.

The ACO enhanced fuzzy rough selection function for web page classification was proposed by Jensen and Shen [31]. Weighting of terms derived from web pages by the TF-IDF weighting scheme. Each sub-set selected by each ant is evaluated by a fluid-based measure of the selected subset in the proposed ACO based characteristic selection algorithm. This measurement and the length of the chosen subset updates the pheromone values. The web pages are then listed after choosing the best feature set. The experiments take place on a small data set containing 280 web pages compiled from Yahoo directory categories Arts and Humanities, Entertainment, computing and internet and the fact that ACO's range of features reduces the area in the highest degree with minimum data loss.

JanakiMeena et al. [32] used ACO for feature selection as well as naive bays in the 20 Newsgroup data set for classification. A ratio of frequency observed with the predicted term is added to the characteristics derived by the bag of terms system as a heuristic test. For parallelization, Map reduction is used. 500 ants and 150 iterations are conducted. Recall and accuracy values are 0.94 and 0.68, respectively, for talk.politics.mideast datasets.

The WebKB dataset has been analyzed by Mangai et al.[33]. Features with minimal variance, knowledge gain and TF-IDF methods of Ward will be picked. The minimum variance measurement of Ward is first used to recognize the redundant function clusters on a web page. Each cluster retains the best illustrative features and eliminates the others. Removal of these redundant characteristics helps reduce the use of resources during classification. The features of these clusters, then kNN, SVM, Naive Bays and C4, are selected following the clustering phase. For classification, 5 classifiers are used with 10 times cross-validation. Good instances are used in the course web pages and negative instances are the student web pages. The feature selection method proposed is contrasted with other similar methods of feature selection. Experiments have shown that the approach suggested is better at minimizing both the number of functions and the classifier training time, than most other feature selection approaches. With kNN and SVM classifiers correspondingly 95.00% and 95.65% accuracy values are attained.

www.turkijphysiotherrehabil.org
2.1. Problem Definition

The categorization of the website is separated by the results of the analyses into several subtitles. Below are some of the headings:

Classification of the main subject: Subject-based classification of the Web page. Art, business, sports, for example

Functional classification: Classification by functional mechanism of the web page. For instance, homepage, login page, administration page.

Emotion classification: Classification made in a certain way to understand the viewpoint of the author. The application has been established according to the key topics in this analysis.

Text Classification and web page problems have comparisons, but they vary from one another structurally.

Firstly, highly structural data sources that are well ordered, semantically and structurelously linked between phrases and paragraphs are the documents listed in the classical text classification difficulties. The strategy of the text is fully controlled by the author in these kinds of documents. This is why those special features can be found in prose. This allows the creation of applications such as the authorship of the text. The structure on web pages is diverse from other sources of text data. The web page, HTML contents, tags, photos, audio files and videos can be used. Text contents may not be related to each other in semanticized or structural terms. On the same website you can find articles on entirely different subjects.

Textual material, including section headings, may not be a full sentence on the web page. The websites which mainly contain textual content and may contain visual elements in their entirety. Restricted textual information and a broader range of types of data which need to be examined in the web page classification problem in comparison with the classification of traditional textual documents. Secondly, web pages have tag component HTML content. The users are visually presented with these HTML materials. A web page has a raw output and a visual output, unlike traditional text documents (browser screen). Links between websites and other web pages or documents are finally created. When classifying a website, it is very interesting to examine hypertexts. Any information on the web page analyzed can be found by links to other sources. This knowledge may be useful for the website classification. There are the variations between classification of text and classification of websites.

III. PROPOSED METHODOLOGY

Instead of choosing functions one by one, NBW-MHO is used to choose a range of functions from the broad feature space extracted from our websites. Our selection method selects characteristics as function classes. Since it is not enough to use a single function to decide the website class, NBW-MHO increases the run-time of all pheromone update computations, including one-by-one features to each spider's selected feature list. However, if we select a number of features, for each selected set we only carry out the required calculations once. For the webpage classification we use the NBW-MHO pheromone update format so that our spiders are not blind; we feed them with each term's term frequency. So, before choosing features, you have an idea about words. We examined the effects on the classification output using URL attributes, tags, tagged words and a bag of terms. We used larger datasets with larger functional spaces. We also examined which tags are more significant to classify web pages. The working flow of the planned manner is shown in figure 1.
3.1 Feature Extraction

All the `<title>`, `<h1>`, `<h2>`, `<h3>`, `<a>`, `<b>`, `<em>`, `<strong>`, `<li>`, and `<p>` tags, text information, and web pages URLs are used during the feature removal process. All words and URLs of the respective Web pages in the training package from each of the tags listed above are taken. The elimination of the word and the stemming algorithm of Porter are implemented after the term extraction. Each terminal and its tag or URL pair is a functionality. In `<title>`, in `<li>` tags and in the URL, for example, a word "program" is taken as a different functionality, and is called "tagged words" method by this function extraction process. For instance `<strong>`, `<b>`, `<em>` and `<i>` terms from similar HTML tags are grouped to reduce function space.

Features of four different function sets are chosen for each class in this analysis. Functionality is only derived from the URLs of web pages in the first package. Secondly, for extraction of functionality only `<title>` tags are used. All words which are listed in the web pages irrespective of their HTML tag are used as features in the third feature removal process. In other words, the phrase in the text is used irrespective of its location. The "bag-of-terms" method is this extraction technique. Finally, all words mentioned above in each of the HTML tags are used for functionality. That is to say, a phrase in various HTML tags is considered a different characteristic.

Depending on the data set and the extraction process, the number of features can vary. Table 1 shows the number of features for each class of all data sets regarding the extraction method used.

<table>
<thead>
<tr>
<th>Class</th>
<th>Feature extraction method</th>
<th>Tagged terms</th>
<th>Bag of terms</th>
<th><code>&lt;title&gt;</code> tag</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course</td>
<td>Tagged terms</td>
<td>33519</td>
<td>16344</td>
<td>305</td>
<td>479</td>
</tr>
<tr>
<td>Project</td>
<td>Bag of terms</td>
<td>30855</td>
<td>15307</td>
<td>596</td>
<td>686</td>
</tr>
<tr>
<td>Student</td>
<td><code>&lt;title&gt;</code> tag</td>
<td>49453</td>
<td>22245</td>
<td>1985</td>
<td>1554</td>
</tr>
<tr>
<td>Faculty</td>
<td>URL</td>
<td>47376</td>
<td>24641</td>
<td>1502</td>
<td>1208</td>
</tr>
<tr>
<td>Conference</td>
<td></td>
<td>34951</td>
<td>18572</td>
<td>890</td>
<td>1115</td>
</tr>
</tbody>
</table>

For example, when the tagged terms method is used for the courses class, 33519 features is extracted. If the tag is taken into account, the number of extracted features for this class decreases to 305.

3.2 Novel Black Widow meta-heuristic Optimization (NBW-MHO)
Like any other algorithm, the proposed algorithm can develop with the initial population of spiders, so each spider is a solution. This first pair of spiders is trying to recreate a novel generation. The black widow eats the man during or after mat. He then replaces the semen kept in his sperm with egg yolk. After 11 days the spiders lay eggs. They get pregnant numerous days a week. During this time the brothers become cannibals. Then they are carried through the air.

a. Initial population

To resolve an optimization problematic, the values of the difficult must be formulated as an appropriate construction to solve the present problem. In black widow meta-heuristic optimization terminology, this structure is referred to as "chromosome" or "particle position", but here in the BWO algorithm, "widow". In the BWO algorithm, the Black Widow Spider is seen as a possible solution to any problem. Every one black widow spider represents the value of a variable problem. In this article, this structure should be seen as an array for performing test functions.

In a $N_{var}$-dimensional optimization problem, a widow is an array of $1 \times N_{var}$ r signifying the problem to solution. This array is distinct as shadows: Widow= $[x_1, x_2, ..., x_{N_{var}}]$. Every adjustable values ($x_1, x_2, ..., x_{N_{var}}$) is floating-point numeral. The fitness of widow is reached by evaluation of $(x_1, x_2, ..., x_{N_{var}})$ fitness function. So Fitness = (widow) =$(x_1, x_2, ..., x_{N_{var}})$.

To run the optimization algorithm, an entrant widow matrix of size $N_{pop} \times N_{var}$ is generated with an initial spider population. Parent pairs are then randomly selected to conduct the mating stage.

b. Procreate

As the sets of each other, they begin to mate to breed a fresh generation. At the same time, each pair mates separately from the others in their network. In the real world, about 1,000 eggs are produced each time they mate, but in the end some of the stronger spider chicks survived. Well, here, in this algorithm, a matrix called Alpha must be created to reproduce when the widow's matrix contains random numbers. Then, the offspring are created using $\alpha$ with the following equation (Equation 1), where $x_1$ and $x_2$ are parents, $y_1$ and $y_2$ are children.

$$\begin{align*}
y_1 &= \alpha \times x_1 + (1 - \alpha) \times x_2 \\
y_2 &= \alpha \times x_2 + (1 - \alpha) \times x_1
\end{align*}$$

(1)

This process is repetitive $N_{var}/2$ times and there is no need to duplicate randomly selected numbers. In the end, children and mothers are additional to the ranks and arranged by fitness scores. Based on the cannibal rating, some of the best people have now been added to the afresh formed population. These measures relate to all couples.

c. Cannibalism

Here we have three types of male predators. The initial is sexual racism, in which a black widow eats her partner during or after intercourse. With these algorithms, we were able to identify women and men based on their fitness values. The second kind is related to cannibals, in which strong spiders eat their pathetic allies. In these algorithms we set the Cannibalism Score (CR), which determines the number of survivors. In particular circumstances, a third species of ogre is often found, with small spiders eating their mother. We use fitness values to identify weak or strong spiders.

d. Mutation

At this point, we arbitrarily select Mute Pop individuals from the population. As shown in Figure 2, each of the selected solutions arbitrarily changes two basics in the array. Intended the mute pop based on the mutation rate.

![Figure 2. Mutation.](https://www.turkphysiotherrehabil.org)
e. **Convergence**

Similar to other algorithms, three closing state of affairs can be deliberated: (a) a predetermined sum of repetitions. (B) The best widow compliance rating for most delegates does not change. (C) Specified accuracy. The pseudocode shown in the figure summarizes the main phases of BWO. The next section discusses some of the issues with optimizing the BWO test using. Since the best solutions for testing activities are already known, the availability of a certain level of accuracy is taken into account to determine the level of accuracy of the experimental algorithm. Also, Section 4 sets the maximum repetitions in the experiments as stop conditions.

f. **Parameter setting**

The projected scheme has some parameters that are required to get the best results. These factors comprise purchase rate (PP), cannibalism rate (CR) and conversion rate (PM). The parameters should be adjusted accordingly so that the algorithm can find the best solutions. Better handling of a number of parameters, higher navigation capability on any local um platform and higher ability to locate globally. Therefore, the exact number of parameters can allow to control the balance between the absorption phase and the inspection phase. The proposed algorithm has three important control factors, including PP, CR and PM. PP is the ratio of ownership that regulates how many people are involved in the product. By regulatory the production of different offspring, this parameter delivers more variety and more possibilities for a clear definition of research location. The control parameter of the CR cannibal operator is to exclude unqualified people from the population. The Prime Minister is the percentage of people who have changed. The fair value of these parameters can strike a balance between use and search. This factor can control the migration of agents in the local phase from global and lead them to better resolution.

3.3 **Classification**

It is clear that two components - forward propagation and rear spreading - form the Recursive Neural Network (RNN) model. Advanced spreads are responsible for measuring the output values and Back spreads the residues accumulated to upgrade weights that are not central to the regular training of the neural network. Figure 3 illustrates the RNN model block diagram.

![Figure 3: RNN models Semantic tree diagram.](image)

The standard RNN is used for solving inductive inference tasks in complex symbolic structures of arbitrary size as a classical neural network context (such as logical terms, trees, or graphs). When a sentence is presented the RNN will analyze it and calculate the vector of each word in a binary semantic tree. The RNN calculates parent vectors in a bottom-up way during the forward-propagation training phase. This is the composition equation:

\[ p_1 = f(W[c_2] + b), p_2 = f(W[c_1] + b) \]  

(2)
Where $f$ is the activation function; $W \in \mathbb{R}^{d \times 2d}$ is the weight matrix, where $d$ is the dimensionality of the vector; and $b$ is the bias. In order to calculate its label probabilities, each parent vector $p_i$ shall be given as an element to a softmax classifier as specified in Eq.2:

$$y^p = \text{softmax}(W_e, p)$$  (3)

Where $W_e \in \mathbb{R}^{3 \times d}$ is the classification matrix. In this recursive process, the vector and classifying result of the node will gradually converge. After the vector of the leaf node is given, the RNN can ultimately map the semantic illustration of the entire tree into the root vector.

IV. RESULTS AND DISCUSSION

In this section, validation of proposed methodology is presented along with experimental setup, performance metrics and description of dataset. The explanation of each section is given as follows:

4.1 Experimental Setup

In this study Perl was used as the script language for the function extraction stage and in the Java programming language in the Eclipse environment our proposed feature selection algorithm was implemented. The approach proposed was reviewed with the operating system 7 rating. The hardware used in the experiments includes 16 GB of RAM and a 3.30GHz processor for Intel Xenon E5-2643. The WebKB datasets test our selection method of functionality. The proposed method for selecting the function is used for 250 iterations since we have seen no change in classification efficiency after 250 iterations.

4.2 Dataset Description

In this proposed research analysis, WebKB Dataset is used. WebKB dataset [34] is a collection of web pages which have been compiled from The 4 Universities Dataset Home Page and collected from the World Wide Knowledge Base (Web->KB) of the CMU [35] text learning community. This website is compiled in 1997 by various university computer science departments and manually divided into seven courses: student, staff, personnel, department, course, project etc. The list includes web pages for each class from four universities: Cornell, Texas, Washington, Wisconsin and other several universities. The 8282 websites are divided manually into seven categories, which means that there are 1641 pages in a student's category, 1124 in a faculty, 137 staff, 182 in the department, 932 in a class, 504 in a project and 3764 in other areas. The other class is a list of pages not considered to be the "primary page" and not an instance in the earlier six classes. The webkb dataset contains 867 Cornell University Web pages, 827 Texas University pages, 1205, 1263 Wisconsin University pages and, ultimately, 4120 other universities pages.

This research includes classes of the WebKB dataset, the faculty, the student and the course. Since there are few positive examples in staff and department classes, they are not taken into account. As defined in the website of the WebKB project, training and test data sets are built[36]. For each class, the training set contains appropriate pages belonging to three universities randomly selected and another dataset class. In the test step the fourth pages of the university are included. The training collection comprises of 75% of the unrelated pages from other classes and the other 25% are used in the test set. Table 2 shows the number of web pages used on the rail and test section of the WebKB dataset. For eg, 846 relevant pages and 2822 irrelevant ones for the training and 86 relevant and 942 irrespective of the test phase are included in the course class.

<table>
<thead>
<tr>
<th>Class</th>
<th>Test Relevant/non-relevant</th>
<th>Train Relevant/non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>43/942</td>
<td>1485/2822</td>
</tr>
<tr>
<td>Faculty</td>
<td>42/942</td>
<td>1084/2822</td>
</tr>
<tr>
<td>Course</td>
<td>86/942</td>
<td>846/2822</td>
</tr>
<tr>
<td>Project</td>
<td>26/942</td>
<td>840/2822</td>
</tr>
</tbody>
</table>

4.3 Performance Metrics
One of the key steps is to assess the accuracy of the master classifiers in the field of data extraction and information recovery. Two popular methods are used to estimate the accuracy of a classifier obtained: the error rate and f measurement to estimate the classifier's ability to find the correct category or class of invisible instances. Error rate shows the share of erroneously labeled test set instances. Let X be a test set containing n instances, and m shows a C classification model of the number of erroneous or erroneous instances. C's accuracy can be determined as follows for the right groups of X instances:

\[
\text{Accuracy}(C) = \frac{m}{n}
\]  

(4)

Error rate method omits the substantial expense of machine learning for inaccurate prediction. This problem is largely solved by using F-measure. To determine the value of the F-measure are used two key measures known as precision and recall. Assume that a group of texts belong to a specific class or category S in a test collection. For each test text document, a category mark shall be assigned by the learning classifier. For S, these predictions will be one of the four following groups:

- True positives is signified as (TP)
- True negatives is signified as (TN)
- False positives is signified as (FP)
- False negatives is signified as (FN)

Accuracy is classified as the percentage of Category S text correctly forecasted, while recall is known as the percentage of category S true text documents correctly predicted. F-measuring can be determined as in Eq.(5-7) according to the precision and recall values.

\[
\text{Precision} = \frac{|TP|}{|TP| + |FP|}
\]  

(5)

\[
\text{Recall} = \frac{|TP|}{|TP| + |FN|}
\]  

(6)

\[
F\text{-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

(7)

### 4.4 Performance Analysis in terms of Proposed Feature Selection Technique

In this section, the performance of the proposed NBW-MHO technique is compared with existing feature selection techniques namely Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), Artificial Bee Colony (ABC) and traditional FOA in terms of accuracy, precision, recall and F-measure. Table 3 presents the performance of proposed feature selection technique.

<table>
<thead>
<tr>
<th>Feature Selection Methodology</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSO</td>
<td>87.89</td>
<td>79.12</td>
<td>80.92</td>
<td>85.27</td>
</tr>
<tr>
<td>WOA</td>
<td>72.30</td>
<td>72.50</td>
<td>73.69</td>
<td>73.07</td>
</tr>
<tr>
<td>ABC</td>
<td>81.25</td>
<td>65.07</td>
<td>88.06</td>
<td>69.28</td>
</tr>
<tr>
<td>FOA</td>
<td>77.26</td>
<td>92.04</td>
<td>93.17</td>
<td>94.08</td>
</tr>
<tr>
<td>Proposed NBW-MHO</td>
<td>95.20</td>
<td>97.64</td>
<td>98.20</td>
<td>98.67</td>
</tr>
</tbody>
</table>
From the Table 3 and Figure 4, it is clearly shows that the proposed NBW-MHO achieved better performance than existing PSO, WOA, ABC and traditional FOA. The proposed NBW-MHO achieved nearly 95% to 98% on all metrics includes accuracy, precision, recall and F-measure.

### 4.5 Performance Analysis in terms of proposed classifiers

In this section, the performance of the proposed three classifiers such as RNN is validated with existing techniques namely CNN and recursive neural network is validated with and without feature selection technique called NBW-MHO in terms of all parameters, which is given in Table 4.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Classifiers</th>
<th>Parameter Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>Precision (%)</td>
</tr>
<tr>
<td>Without NBW-MHO</td>
<td>CNN</td>
<td>88.14</td>
</tr>
<tr>
<td></td>
<td>Recursive neural network</td>
<td>88.90</td>
</tr>
<tr>
<td></td>
<td>Proposed RNN</td>
<td>92.25</td>
</tr>
<tr>
<td>With NBW-MHO</td>
<td>CNN</td>
<td>92.47</td>
</tr>
<tr>
<td></td>
<td>Recursive neural network</td>
<td>94.80</td>
</tr>
<tr>
<td></td>
<td>Proposed RNN</td>
<td>97.57</td>
</tr>
</tbody>
</table>

The following Figure 5 represents the graphical results of proposed classifier without NBW-MHO in terms of all parameters.
The following Figure 6 represents the graphical results of proposed classifier with NBW-MHO in terms of all parameters.

From the comparative study of feature selection with proposed classifier, it is clearly stated that the proposed RNN classifier achieved better results only with proposed NBW-MHO selection technique than CNN and recursive neural network.

V. CONCLUSION

As web pages expand quickly each day, it makes it difficult for a web crawler to read and organize web pages. This problem increases daily importance of the web classification process. We have implemented our algorithm with standard data sets in this paper, and the results show that our algorithm is higher than current algorithms. In this paper we have put forward a feature selection algorithm for optimization in order to suggest top N features to the ML model for the prediction of the correct web document class. Web pages of a corpus have links that bind each other; these links provide an additional benefit compared to the usual text classification paper. If a website has several links to a series of web pages, the links reflect the link information, and the relation between them is high, which means that the target group is also large. This knowledge is used for classification purposes with the proposed feature selection procedure in the paper. In addition to the ties, the approach proposed considers the side details to improve the accuracy of classification. The results have shown that our classification model has
encouraging positive effects on web documents. We checked our classification model with standard benchmark data. Information ontology and traffic relations will be considered as an additional parameter in the future.

REFERENCE