Received: May 12, 2022. Revised: June 17, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Classification of Web Pages Using the Machine Learning Algorithms with Web Page Recommendations

Chaithra[1]* Lingaraju Gowdru Malleshappa[2] Jagannatha Sreenivasaiah[3]

[1]*Department of Computer Science &Engineering,*
Sapthagiri College of Engineering, Hesarghatta Road, Bangalore, India
[2] *Department of Information Science and Engineering,*
East Point College of Engineering and Technology, Bangalore, India
[3] *Department of Computer Applications, MSRIT, MSRIT Nagar, Bangalore, India*
* Corresponding author's Email: chaithra81@gmail.com

Abstract: The World Wide Web holds a huge source of a variety of information. Web users are provided with a lot of information with numerous options and choices because of which decision-making will be difficult for the users. The users must be provided with a recommendation system that makes it easy for making their decisions, one such system is the web page recommendation system. The problem faced by web users is they can't get the required web pages when they search the web. To provide the solution to this problem web page recommendation systems are proposed. Recommendation system provides the users with interesting Webpages or websites where they can be reduced their searching time or surfing time. The web user needs effective and some suggestions for accessing the website efficiently. Therefore, the web recommendation systems are very useful for the user by efficiently handling the website. In the proposed work Web page recommendation system is implemented based on web page classification and by determining the page rank, where the classification of the web pages is done based on the machine learning algorithms. The classification techniques are used with the various machine learning algorithms in our work which are k-nearest neighbor, support vector machine, ADABOOSTER, and entropy-based ensemble Random Forest. The proposed algorithm in our research work is an entropy-based ensemble random forest. The novelty of this research work is the use of the entropy-based ensemble random forest. The web scraping technique has been used to fetch images and text from various websites. The data collected of 3000 are pre-processed and applied to the various classifiers to classify the web pages. The web pages were classified with the highest accuracy of 99.55% using the entropy-based ensemble random forest. The comparison of the entropy based random forest and the Gini-based random forest is shown to achieve the novelty. We have taken the data from the web pages in the form of images of around 17034 are classified them using the convolution neural network and pre-trained convolution neural network known as Resnet50. Compared its results with our proposed ensemble entropy-based random forest. Based on the web classification the web page recommendation is done. The page ranks for the collected web pages are calculated and the web page which will be similar to the given web page will be given as output. The page rank algorithm is used for the calculation of page rank. For every source website, a similar target web page is calculated. This is going to form the recommendation system.

Keywords: Web page recommendation system, Web page classification, Entropy-based ensemble random forest.

1. Introduction

The web is a huge source of information, classifying the web page according to the user's needs can be achieved by using various machine learning algorithms. The web structures are complicated and it's very large. The users may get confused about unnecessary information. It is required to improve the surfing experience and users should be provided with the required information in less time. The web page recommendation system is the only solution. The classification of the web pages is by using the algorithms of machine learning like a k-nearest neighbor, support vector machine,

AdaBooster, and proposed entropy-based ensemble random forest are used. The web page's text and images are taken from the various web pages by using the web scraping technique. The text and images which are collected are pre-processed & classified and the highest accuracy obtained is 99.55%. The web pages data in the form of images around 17034 and classified it using the CNN and obtained an accuracy of 99%. The webpage recommendation system proposed identifies the similarity of web pages according to the user's interest and makes the information available to the users. The web page recommendation system recommends more preferable websites to web users. In the search engine, the web user types the keyword to search vast information of web pages on the world wide web. Therefore, the researchers developed efficient statistical classification or machine learning techniques namely SVM, K-nearest neighbor (KNN) classifiers, Bayesian classifiers, and neural networks to identify the web pages by the search engine in an effective way. Web pages are ranked [1] using the page ranking algorithm to find the keyword of interest to users. The source web page of the user's interest is taken and the target web page is searched.

The overall organization of the paper is given as follows: Section 2 represents the literature survey about the web page classification. The various algorithms used for the feature selection and web page classifications are discussed in this section. A detailed explanation of the proposed web classification method & page rank algorithm is described in section 3. Section 4 presents the results and discussion of the proposed method. Finally, the conclusion is made in section 5

2. Literature review

Raju, V., & Srinivasan, N. [2] have explained the web page recommendation system, in the proposed system data has been categorized into potential and non-potential by use of an algorithm known as Levenberg–Marquardt firefly neural network algorithm, and the prediction is done by another clustering algorithm known as K-means algorithm. The recommendation accuracy is estimated corresponding to the various iterations. To take potential data for clustering, this phase is performed with the aid of the IFCM technique. Prasanth, A.[3] has told in his work that because of the rapid development of the world wide web the web users have many choices and it has become difficult for the web users in making decisions, because of which the recommendation system is suggested for accessing the web site effectively. Recommendation results are very helpful in handling the website efficiently. This work predicts the user navigational preferences from previous activities used to recommend the websites to web users. A collaborative filtering mechanism has been used to analyse one user's interest with another user's interest. The Chameleon hierarchical clustering process is used for discovering the frequent search pattern. Chu, W. T., & Tsai, Y. L.[4] in this work, two recommendation approaches are taken one is content-based filtering & the other one is collaborative filtering for predicting a favorite restaurant. The filtering method is effective for information that is in visual form. The performance is taken as the overall performances of MF & BPRMF are compared with different approaches. The improvement of the BPRMF approach by integrating visual information more effectively can be future work.

Katarya, R., & Verma, O. P [5] in this work web page recommendation is has been utilized for the personalized web page recommendations to predict the ratings for the items, the web page recommendation system majority does not support the information's which are sequential on web pages, instead mainly focuses on the content information. In this work, sequential information from web pages is used in the recommendation system. The top N clusters were received when the Fuzzy C mean clustering algorithm was applied to it. Similarity among the users and target users is determined & web page weight is evaluated. Web page recommender system problems are solved by providing the users next web page visit. For the experimental results, the real dataset of MNSBC is used. A 5000 user entry, 6 entries per user was there in the dataset. The accuracy obtained is 33%. The future work will be focused on inclusion of privacy, trust and social networks with the utilization of hybrid intelligent systems. Wagh, R., & Patil, J. [9] proposed a novel web page recommender system to provide a good experience to web users. In this work web, the usage mining technique is used for the web pages. Preprocessing, analysis of the web pages relationship, clustering, and classification of data are performed. The threshold value is used to recommend the web pages. Two experiments are conducted the first experiment is modeling & clustering of web pages is done using an enhanced depth-first search algorithm, the second experiment LCS algorithm is used to classify user active sessions in one of the clusters. The novel measures of finding the relationship, use of threshold values at the time of formation of clusters as well as at the time of recommendation of web pages give us better results in terms of improved visit coherence,

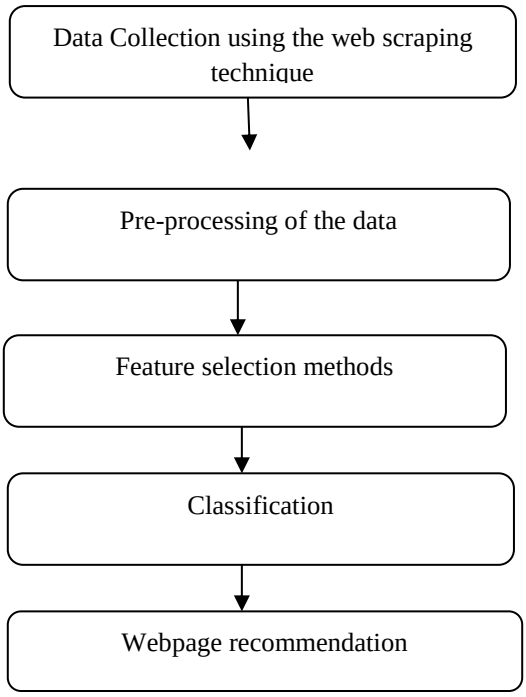

Figure. 1 Proposed method of Web page classification and Recommendation

accuracy, coverage, and F1 measures. Accuracy of 61% was used for the web page recommendation. Still, there is scope for the increase of accuracy.

Nilashi, M., Ibrahim, O., & Bagherifard, K. [15] in this work a new hybrid recommendation method is proposed based on the collaborative filtering (CF) approaches. Sparsity and scalability are the two main drawbacks that are solved by recommender systems. It uses dimensionality reduction and ontology techniques. Items most similar are determined by using singular value decomposition (SVD). In future scalability issue of CF can be solved by incorporating the incremental SVD into the prediction models.

3. Proposed methodology

The proposed method of web page classification and recommendation is explained in Fig. 1 the data is collected from the web using the web scraping technique, and feature selection methods are applied. Fig. 1 shows the proposed EERF classifier used to classify the dataset. Once classification is achieved, the page rank algorithm is applied to achieve the web page recommendation. The methodologies are explained in detail in the following section.

3.1 Methodology to classify web page usingproposed EERF method

Data collection: The text & image data are extracted from type classes like animal, food & flower. The scraping technique is used for the extraction of the data stored in the hypertext mark-up language format. The python library called BeautifulSoup is used to fetch the data from the HTML files. The image and text data are scraped using BeautifulSoup packages and stored in a structured format for further processing.

Pre-processing of the data: During the pre-processing of the text includes the removal of special characters, upper cases, lower cases, and tokenization and in a similar way the image pre-processing includes the normalization. Classification: The proposed EERF classifier is implemented in this study to classify the classes of the collected image and text data only after attaining the feature values. The random forest is a superior classification technique for training the large dataset because it uses only a low number of feature values to classify the three different classes animals, flowers, and food. Web page recommendation: Data collected from Webpages URLs are collected and used for the web page recommendation by using the page rank algorithm.

3.1.1. Scraping techniques

The process of retrieving information from different websites is known as scraping activity. Moreover, the scraping process can be manually or automatically carried out for information retrieval. The scraping techniques are normally categorized into three types web scraping, screen scraping, and report mining. In screen scraping, the user's screen data is captured from one of the applications and is transmitted to another application to display the same information. Report mining is defined as the process of extracting the information from computer reports, where the extracted information is in a human-readable format. At last, the necessary information retrieved from the different websites is transferred into a structured form for further analysis. In this study, web scraping is used to classify the collected data from web pages is explained in the following sections.

3.1.2. Web scraping techniques

The various unstructured information extracted from different websites using web scraping techniques is converted into a structured form for further analytical processes. Any remote servers or local computers of users can be used to store those structural data. Recently, many researchers have used web scraping techniques to create their own data sets related to the extracted information from journal

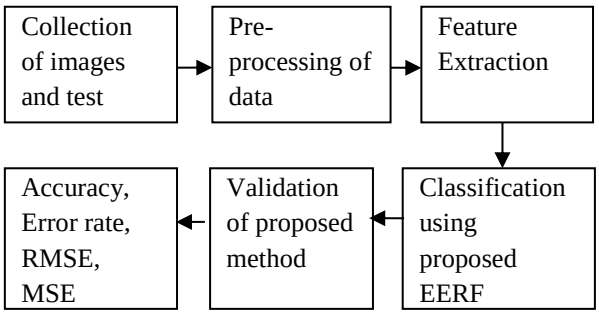

Figure. 2 Working flow of Proposed EERF method

articles and text mining projects. The web scraping software automatically extracts the important information from multiple pages loaded on the websites to avoid misinterpretation and omission of data. According to a user's query, the data can be extracted from the websites and the user can access it at any time.

3.1.3. Extraction of unstructured content

The necessary data fetched from websites is called web data extraction, which converts the semi-structured and unstructured data into a structured format and is stored in a database or warehouse. Moreover, a vast amount of unstructured data is present on the Internet. The unstructured data only become useful for the analytical process when it is framed into a unique form called structured data.

3.1.4. Tools of web scraping

Web-scraping tools are one of the most available open-source tools online. The tools for web scraping frameworks' effectively parse the websites and extract required content from websites. Some of the important scraping tools are cURL, OpenRefine, Wget, and Web scraping service. The data and webpages can be retrieved through an application programming interface (API) using cURL and Wget, where the process of collecting the data uses the web scraping service tools and the cleaning process uses the OpenRefine tools. The researchers use other web scraping tools such as Mozenda, web content extractor, visual web ripper, scrappy, and import.io to extract the required information from web pages.

In this study, the web contents are retrieved using an effective web scraping method from websites, and the data collected are converted into a form of structure for analytical processes. The collected data are trained and the performance of the proposed EERF method on categorizing the documents is validated using the proposed EERF classifier Fig. 2 presents the working procedure of the proposed EERF method.

3.1.5. Image and text content scraping

The text and image contents are extracted from three different classes: animal, food, and flowers. The scraping protocol is used to access the target websites and those data are stored in hypertext mark-up language (HTML) format using different objects. Next, the poor-quality images constructed by HTML are eliminated from the extracted data, and only the useful information is separated for final classification. The available python libraries are used to extract the information from the HTML pages. In this research, a package or library called BeautifulSoup is used to fetch the data from the HTML files. The image and text data are scraped using BeautifulSoup packages and stored in a structured format for further processing.

3.1.6. Pre-Processing of text and image data

The extracted features from three different classes of animals, food, and flowers are pre-processed to improve the overall performance of the EERF classifier. The pre-processing of the text includes the removal of special characters, upper cases, lower cases, and tokenization and similarly, the image pre-processing includes the normalization. In a text, individual pieces are obtained by dividing the words using tokenization and every word is called a token. A statement with punctuations and complex representation is normalized by tokenization, where every collected text is converted into tokens in this phase. The pixel intensity values of collected images are changed using the normalization technique. The pre-processed data are given as input to extract the features of three classes, namely animals, flowers, and food. Here, the features are represented as the numerical values of those three classes.

3.1.7. Feature extraction of image and text data

The color distribution of an image is represented using a color histogram, which is used to extract the features from the collected information or data. An appropriate range is achieved by dividing the space of images, where this range is arranged as a regular grid and contains the same color values. The statistical distributions of colors are represented by a color histogram also known as smooth function, which is an essential tone of the image. The total number of various color type proportions with spatial location is focused only on the color histogram of the images. The texts are extracted from the input collected data using the label encoder techniques, which convert the extracted text data into numerical

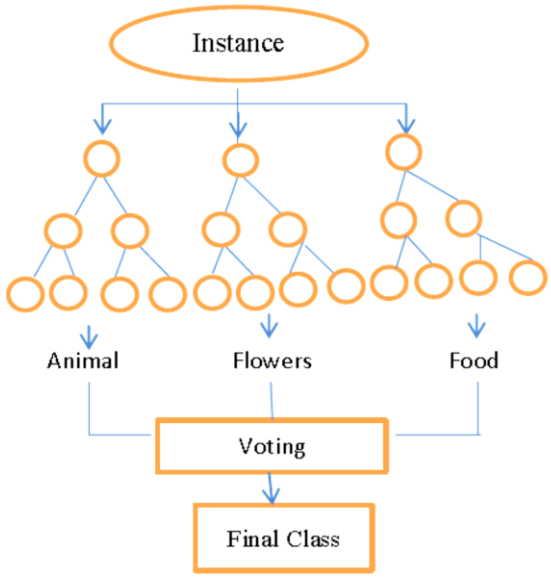

Figure. 3 Graphical representation of proposed random forest

values for further processing.

3.1.8. Classification using entropy-based ensemble random forest classifier

The proposed EERF classifier is implemented in this study to classify the classes of the collected image and text data only after attaining the feature values. The random forest is a superior classification technique for training the large dataset because it uses only a low number of feature values to classify the three different classes animal, flowers, and food. Additionally, the RF classification technique is a non-parametric pattern technique that significantly diminishes the issue of probability density complexity. In the RF classification technique, each tree is assumed as distinct classifiers, which are utilized to achieve better decision-making. The graphical illustration of the proposed random forest classifier is presented in Fig. 3.

The ensemble classification technique (random forest) works based on the principle of bagging, which uses a decision tree as a base classification technique. Initially, extracted feature vectors are randomly sampled for training sets N. Next, sub-feature values are selected from the extracted feature vectors if $m(m < M)$, where M is indicated as extracted feature vectors. Further, select m feature vectors from the M feature values and then split the nodes of the tree using best spilled on the m dimensional feature vectors.

The error rate of the proposed random forest classifier depends on two factors, which are given as follows:

- Tree strength should be high to diminish the error rate.
- Correlation among trees needs to be below to reduce the error rate.

However, the proposed random forest is insufficient to measure the quality of information splits in the tree structure. To address the aforementioned issue, this study implemented the entropy function to measure the information split's quality. The mathematical Eq. (1) shows the criterion entropy function, which is used in this study to measure the quality of information splits.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (1)$$

Where S is an entropy value and p_i have described the probability of an element or class"i" in the collected data. The entropy function is used in the random forest to develop the proposed EERF classifier and the following section discusses the validation of the proposed EERF method. Based on the entropy the average information of that attribute is calculated, finally the information gain is calculated using the following formulas:-
Average information-

$$\text{I(Attribute)} = \sum \frac{pi}{p} * E(c) \quad (2)$$

$$\text{Gain(Attribute)} = \text{E(S)} - \text{I(Attribute)} \quad (3)$$

Data collection: Image data are collected from the various web pages and pre-processed. A further comparison of proposed ensemble entropy based random forest with traditional CNN and pre-trained model of from Keras transfer learning has been done. Web page recommendation is performed for the data collected from the Webpage's corresponding URLs are collected and used for the web page recommendation by using the page rank algorithm.

3.2 Classification based on convolution neural network (CNN) classifier

Feature extraction utilizing ConvNets, ConvNets at present are the go-to models, for visual acknowledgment. The pre-prepared model is utilized on ImageNet saves time and having a best-in-class model out of the crate is called "transfer learning". The elements the model has learned not the class

Table 1. Model summary of CNN

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 rescaling (Rescaling)       (None, 150, 150, 3)       0

 conv2d (Conv2D)             (None, 150, 150, 32)      896

 max_pooling2d (MaxPooling2D  (None, 75, 75, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 75, 75, 64)        18496

 conv2d_2 (Conv2D)           (None, 75, 75, 64)        36928

 max_pooling2d_1 (MaxPooling  (None, 37, 37, 64)       0
 2D)

 conv2d_3 (Conv2D)           (None, 37, 37, 128)       73856

 conv2d_4 (Conv2D)           (None, 37, 37, 128)       147584

 max_pooling2d_2 (MaxPooling  (None, 18, 18, 128)      0
 2D)

 flatten (Flatten)           (None, 41472)             0

 dense (Dense)               (None, 64)                2654272

 dropout (Dropout)           (None, 64)                0

 dense_1 (Dense)             (None, 6)                 390

=================================================================
Total params: 2,932,422
Trainable params: 2,932,422
Non-trainable params: 0
_________________________________________________________________
```

Table 2. Model summary of pre-trained Resnet50

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
resnet50 (Functional)        (None, 7, 7, 2048)        23587712
_________________________________________________________________
flatten_1 (Flatten)          (None, 100352)            0
_________________________________________________________________
dense_2 (Dense)              (None, 6)                 602118
=================================================================
Total params: 24,189,830
Trainable params: 602,118
Non-trainable params: 23,587,712
_________________________________________________________________
```

probabilities and pre-learned layers recognize shapes, designs, and so on will ideally think of portrayals that outcome in significant proposals. Consequently, the yield layer treats the remainder of the ConvNet as a component extractor for website pages. In the execution of the recommender, the Keras with TensorFlow is used as a back end. For each picture in dataset, save the leveled yield cluster of the last secret layer of the model. The new component cluster works out the quantity of nearest neighbors of an objective picture/banner dependent on the Euclidean distance of the exhibits to one another. The philosophy might have many question page sub-charts that contain similar ideas and relations in various plans. Weight is related to each edge to consider the number of relations connecting ideas in the chosen page based on the arrangement of clarified relations. The likelihood of finding is registered in a specific page a connection among ideas that could be the one important to the client.

The CNN classifier is used to classify the images collected of the size 17034. The accuracy obtained is about 85%. The techniques were required for the data classification to increase the accuracy. The CNN classifies the image with an accuracy of 85%, the model summary is given in Table 1. The pre-trained model from Keras transfer learning is used to improve the accuracy of the CNN classifier to 99%. The model summary description of the CNN model is given in Table 2. Convolutional neural network is a well-known method in the field of computer vision applications. CNN is one of the classes of deep neural networks which are used to analyse images. The CNN architecture is dominant in recognizing objects from a picture as well as in the video. CNN is used in the applications of image recognition, neural language processing, etc.

3.2.1 Architecture of a convolutional neural network

As the first step, an image is given as the input image. The input image goes through several steps which are the convolutional part of the network. Finally, the convolution neural network can predict the image

A pre-trained model was trained on a large dataset for performing image classification on large scale. The pre-trained models Resnet50 use transfer learning to customize the model to a given task, the model summary of pre-trained model Resnet50 is shown in Table 2. Deeper neural networks are more difficult to train so we are using the residual learning framework to ease the training of networks that are substantially deeper than those used previously. The residual networks are easier to optimize and can gain accuracy from considerably increased depth.

The page rank makes the content of the internet valuable and it spreads at a faster rate. A few of the data are downhill acceleration depending on the page evaluation value. The main idea of using the page rank algorithm is to change the page rank value on basis of the time series is being analysed, some of the URLs expected are predicted in future time. This type of prediction forms the effective parameter for the search engines that provides the retrieval.

3.2.2. Quantitative analysis

Here in re-enactment work, right off the bat, info

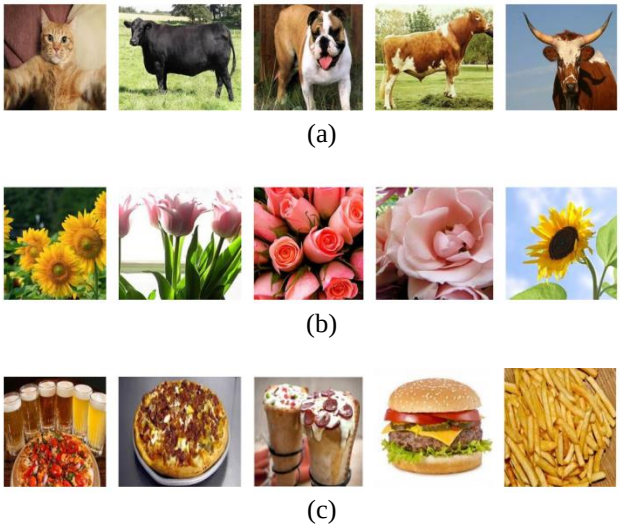

Figure. 4 Sample of collected images: (a) animals; (b) flowers, and (c) food

is taken as a dataset from the website pages, and afterward, pre-handling is performed utilizing a tag-based methodology that is included the header, tail, and body of the substance. Then, the alterations of the client, for example, changing the substance of the website page should likewise be possible with their advancement without the mediation of the server. A while later, the altered website page content is put away in the information base. In addition, the evaluations are given by the clients. Then, at that point, these evaluations are arranged by utilizing CNN classifier and the need for the website page substance is acquired dependent on these appraisals. Ultimately, the client can get to the pages and prescribe them to others dependent on the focus on information. In this part, we present the presentation of the proposed model. The exhibition of the proposed model is assessed and broken down by utilizing different execution measurements like exactness, accuracy, review, and F-score. Here the boundary, for example, page rank exactness, information source precision, the page suggests exactness for discovering website page rank utilized in assessing the exhibition of the proposed model

4. Results and discussion

The simulations are carried out on text and images which are collected to validate the efficiency of the proposed EERF method using accuracy, root mean square error (RMSE), area under curve (AUC), error rate, precision, mean absolute error (MAE), Recall and F-score. The proposed method is implemented using Python 3.7.3, 8GB RAM, and an Intel core i5 processor with 2.2 GHz. The image and text data with three classes' flowers, food, and animals are collected in this research, where each class contains 1000 text and 1000 images. Fig. 4 presents the sample images of collected data for every class.

During the training process, the proposed EERF method uses 70% of collected data, and 30% of data is used for the process of testing. The existing techniques such as KNN and SVM are validated only on either adult content/medical content or collected data for fake news detection, but this research worked on the randomly collected data of animal, flower, and food classes. The results and comparative study of proposed and existing methods are discussed in the following section.

4.1 Parameter metrics

The EERF method is validated using important parameters like F-score, Accuracy, MSE, RMSE, recall, and precision. The precision is used to measure the truly corrected documents, where a portion of corrected documents are identified by the recall. The mathematical Eqs. (4) and (5) show the precision and recall.

$$Precision = \frac{Documents\ assigned\ correctly}{Total\ assigned\ documents} = \frac{A}{A+B} \quad (4)$$

$$Recall = \frac{Documents\ assigned\ correctly}{Documents\ belong\ to\ the\ class} = \frac{A}{A+D} \quad (5)$$

Where A and B are illustrated several documents assigned correctly and incorrectly. Next, the total number of documents rejected incorrectly is represented as D. The error rate defines the ratio of incorrect documents to the total number of documents, whereas accuracy defines the ratio of correctly assigned documents to the total number of documents. Accuracy and error rate are mathematically expressed in the following Eqs. (6) and (7).

$$Accuracy = \frac{correctlyassigneddocuments}{Totalnumberofdocuments} = \frac{A+C}{A+B+C+D} \quad (6)$$

$$Error\ Rate = \frac{Incorrect\ Assignments}{Total\ number\ of\ documents} = \frac{B+D}{A+B+C+D} \quad (7)$$

Table 3. Performance of proposed EERF method in terms of precision, recall and F-measure

Methodology	Precision (%)			Recall (%)			F-Measure (%)		
	Animal	Flowers	Food	Animal	Flowers	Food	Animal	Flowers	Food
KNN	57	81	84	94	23	84	71	36	84
SVM	76	78	90	79	74	91	78	76	90
Adaboost classifier	79	74	89	74	76	92	76	75	91
Proposed EERF	**87**	**86**	**87**	**84**	**82**	**95**	**85**	**84**	**91**

Table 4. Performance of EERF classifier by means of MAE, RMSE and MSE

Methodology	MAE (%)	MSE (%)	RMSE (%)
KNN	37.22	47.22	68.71
SVM	22.55	29.88	54.67
Adaboost classifier	22.44	28.22	53.12
Proposed EERF	**16.00**	**21.55**	**46.42**

Table 5. Overall accuracy, AUC and error rate of proposed EERF classifier

Methodology	Accuracy (%)	Error Rate (%)	AUC (%)
KNN	67.77	32.22	88.7
SVM	81.11	18.88	92.97
AdaBoost classifier	80.44	19.55	93.54
Proposed EERF	**87.0**	**13.22**	**94.60**

Where C is described as the total number of documents rejected correctly. MSE is used to measure the average squared between the estimated and actual values and F-Measure is the harmonic measure of precision and recall, where Eqs. (8), (9), and (10) show the mathematical expression for F-measure, mean square error (MSE) and RMSE.

$$F - measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|^2} \quad (10)$$

Where y is defined as variables in the document and T is represented as time.

4.1.1. Performance analysis of proposed EERF method on image classification

The proposed EERF method performance is evaluated in both collected text and image data and initially, the image data classification using the proposed method is presented in this section.

The existing techniques KNN, SVM, RF, and AdaBoost classifier were implemented on the collected data, and results are taken for the three different classes animals, food, and flowers. Table 3 shows the performance of proposed EERF and existing techniques through precision, recall, and F-measure for all three classes.

In animal class, the existing KNN technique provides poor performance than other techniques using precision, recall, and F-measure, i.e. it achieved only 57% of precision, 94% of recall, and 71 % of F-measure. The KNN technique works based on the distance between the neighbouring nodes, which degrades the overall performance of the KNN technique. The existing SVM technique achieved moderate results nearly 76% to 91% of precision, recall, and F-measure due to the presence of marginal parameters in the SVM. Moreover, the AdaBoost classifier achieved better precision and f-measure, but it provides low accuracy of 80.44% compared to other existing techniques. However, it has failed to improve the quality of information and leads to poor performance in the animal class. The performance of the proposed EERF method has gradually increased due to the addition of the entropy function in the RF classifier. Therefore, the proposed EERF achieved a precision of 87%, recall of 84%, and f-measure of 85% for animal class in image classification.

Table 4 presents the performance of the proposed EERF classifier using RMSE, MSE, and MAE. From

Table 6. Performance analysis of proposed method on three different classes in text classification

Methodology	Precision (%)			Recall (%)			F-Measure (%)		
	Animal	Flowers	Food	Animal	Flowers	Food	Animal	Flowers	Food
KNN	86	66	100	57	90	100	69	76	100
SVM	48	44	95	50	40	100	49	42	98
Adaboost classifier	72	54	99	33	85	100	45	66	99
Proposed EERF	100	98	100	98	100	100	99	99	100

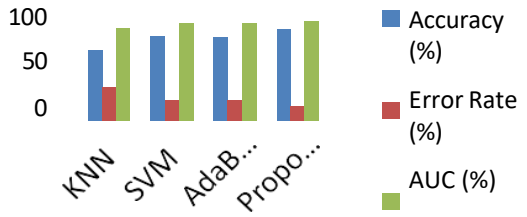

Figure. 5 Performance analysis of proposed method by means of accuracy and AUC

Table 7. Analysis of proposed EERF in terms of MSE, MAE and RMSE

Methodology	MAE (%)	MSE (%)	RMSE (%)
KNN	18.88	19.11	43.71
SVM	39.11	41.33	64.29
Adaboost classifier	29.44	29.44	54.56
Proposed EERF	**0.66**	**0.66**	**8.1**

the value present in Table 4, it is clear that the proposed EERF classifier achieved fewer error rates such as MAE, RMSE, and MSE than other traditional techniques.

Table 5 presents the average accuracy, error rate, and AUC of the proposed EERF method, where the error rate of the proposed EERF method is 13.22%, whereas the existing AdaBoost classifier achieved a 19.55% of error rate. The existing SVM and KNN techniques achieved nearly 20% to 25% of error rate on image data. Among the other existing techniques, the EERF classifier achieved less error rate, due to its effective structure, and the proposed method provides better performance.

Table 5 and Fig. 5 present the validated results of the proposed EERF classifier and other traditional techniques AdaBoost, SVM, and KNN using accuracy and AUC. While compared with other techniques, the AdaBoost classifier provides performance of 80.44% of accuracy and 93.54% of AUC.

The SVM and KNN techniques achieved nearly 76% to 100% of accuracy, whereas the AdaBoost classifier technique achieved 33% to 100% accuracy. The existing techniques SVM and KNN achieved a higher AUC, but the existing techniques failed to provide high-quality images. Therefore, the proposed method introduced the entropy calculation in RF and increased the quality of the collected images, which increases the overall accuracy and AUC. The performance of the proposed EERF method on collected text classification is explained in the following section.

4.1.2 Performance analysis of proposed EERF method on text classification

The result of the experiments conducted on collected text data of three classes is presented in terms of accuracy, error rate, AUC, precision, recall, f-measure, MAE, RMSE, and MSE in this section. The validation of various existing techniques SVM, KNN, and proposed EERF based on the precision, recall, and f-measure of three classes is presented in Table 6.

The existing techniques SVM, KNN, and AdaBoost classifier provide poor performance in precision, recall and F-measure on animal and flower classes, but achieved high performance in food class. For instance, the SVM achieved 44% to 48% of precision in animal and flower classes, whereas the same technique achieved 95% of precision in food classes. The performance of existing techniques is reduced due to the presence of special characters in the texts. The RF classifier effectively pre-processed the text, while collecting the text in various classes and achieved higher performance on precision, recall, and f-measure of all three classes. This research introduced the criterion entropy function in the RF classifier which increased the overall performance of

the proposed EERF classifier. For example, the proposed method achieved 100% of recall, precision, and f-measure on food class.

Table 7 presents the validation results of the proposed EERF using MAE, RMSE, and MSE. The value presents in Table 7 indicates that the existing SVM technique has higher RMSE, MSE, and MAE than other traditional techniques.

Table 8. Performance of proposed EERF in terms of error rate, AUC, and accuracy

Methodology	**Accuracy (%)**	**Error Rate (%)**	**AUC (%)**
KNN	81.22	18.77	99.91
SVM	33.0	67.0	59.97
Adaboost classifier	70.55	29.44	99.67
Proposed EERF	**99.55**	**0.66**	100

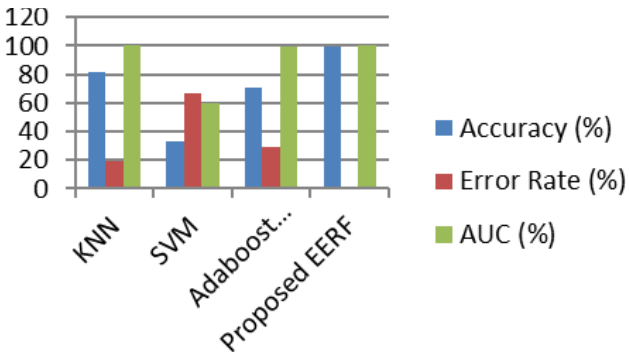

Figure. 6 Graphical illustration of proposed EERF on the by means accuracy and AUC

Table 9. Comparative study of proposed EERF method and existing methods.

Author	**Methodology**	**Input Data Type**	**Accuracy (%)**
Vishwakarma [16]	Rule-Based Classifier	Image	85
	Rule-Based Classifier	Text	86
Ali [17]	SVM with Fuzzy Ontology	Text	97
Proposed Method	EERF classifier	Image	87
Proposed Method	EERF classifier	Text	99.55

The AdaBoost classifier used the most important features of text for classification and achieved nearly 29% of MAE and MSE. The existing KNN achieved nearly 18% to 19% of MSE and MAE. The entropy function in the proposed method classified the text effectively and achieved higher performance, i.e. 0.6% of MSE and MAE. In addition, the overall accuracy, error rate, and AUC of the proposed method are presented in Table 8 and a graphical representation of the overall accuracy and AUC of the proposed EERF classifier is presented in Fig. 6.

While comparing with image data classification it is observed that the error rate of the proposed method on text data classification is highly minimized, i.e. the proposed EERF technique achieved a 0.66 error rate. The SVM technique has the highest error rate (i.e. 41%) of other techniques due to the marginal parameters of SVM. The classification performance of the SVM is degraded by the improper extraction of text features. The experimental result shows that the proposed method effectively classified the text data and achieved less error rate.

The existing KNN and Adaboost classifier achieved nearly 70% to 81% of accuracy, whereas the proposed EERF achieved 100% of AUC in text classification. From the experiments on image and text classification, the validated results proved that the proposed EERF classifier achieves higher performance compared to the traditional techniques.

4.1.3 Comparative study

The proposed method performance is compared with existing techniques, namely, rule-based classifier [16] and SVM with fuzzy ontology [17] using accuracy. Table 9 presents the comparative analysis of the proposed EERF method with the existing fuzzy and rule-based classifier. Table 9 presents comparative results and it proved that the proposed method achieved better performance on image and text data classification. For instance, the existing rule-based classifier achieved only 85% of accuracy, whereas the proposed EERF classifier achieved 87% of accuracy on image data.

Moreover, the rule-based classifier achieved only 86% of accuracy on text data due to the misclassification of text extraction. The existing SVM with fuzzy ontology achieved 97% of accuracy and the performance is degraded due to the massive amount of data. The proposed EERF classifier extracted the text data effectively using a label encoder and achieved the 99.55% of accuracy. The comparative study shows that the proposed EERF method achieved higher performance than the

Table 10. A comparative study of the proposed EERF method and Gini based random forest is shown for image

Methodology (Image)	Precision (%)			Recall (%)			F-Measure (%)			Accuracy
	Animal	Flowers	Food	Animal	Flowers	Food	Animal	Flowers	Food	
Gini Based Random Forest	87	84	88	82	82	95	84	83	92	86
Proposed EERF	87	85	88	84	82	95	85	84	92	87

Table 11. Comparative study of proposed EERF method and Gini based random forest is shown for text classification

Methodology (Text)	Precision (%)			Recall (%)			F-Measure (%)			Accuracy
	Animal	Flowers	Food	Animal	Flowers	Food	Animal	Flowers	Food	
Gini Based Random Forest	100	98	100	98	100	100	99	99	100	99.33
Proposed EERF	100	99	100	99	100	100	99	99	100	99.55

various existing techniques using classification accuracy shown in Table 9. The EERF is compared with Gini based random forest by selecting the n_estimators in both techniques by 1000 to see the performance of both techniques for image classification in Table 10. The EERF is compared with Gini based random forest by selecting the n_estimators in both techniques by 2000 to see the performance of both techniques for text classification in Table 11.

In [21] pre-trained Bert model with deep residual inception model network has been proposed for the classification of the web pages. DMOZ dataset is used where DMOZ dataset is an URL classification dataset from the DMOZ directory, which has 10 different top classes for the classification of web pages into different categories including business, society, science, recreation, shopping, games, arts, business, computers, and health. The same dataset is tested with our proposed model; our proposed model performs very well when compared to the model of [21]. The comparisons are shown in the Table 12.

Table 12. A Comparative study of the proposed EERF method and existing methods.

Author	Methodology	Type of input	Accuracy (%)
Gupta, A. [21]	Pretrained Bert Model with Deep Residual Inception Model Network	Text	66%
Proposed Method	EERF classifier	Text	99.55 %

Table 13. The accuracy of the pre-trained Resnet50 models

Classifier	Accuracy Achieved
CNN	85%
Pre-trained Model Resnet50	99%

4.2 Results of the CNN and pre-trained CNNModels

The convolution neural networks are the type of neural network model. The ability to automatically learn a very huge number of filters which are in parallel to training the dataset of specific problems is associated with this model. The highly specific features can be detected in the input images. The linear operation is performed which involves the multiplication of a set of weights with the input. Similarly, the comparison of CNN with trained Keras accuracy compared with the existing models is shown in Table 13.

The classified data taken from the web pages from both the methods are taken and applied to the page rank algorithm. In the page rank algorithms calculate the web pages page rank using the URL of the pages. For classified groups, we are applying a page rank vector on which page to which link has been connected we are going to calculate, the most recommended page rank. Page vector is selected for recommending the next page, for all the combinations of pages, we are calculating a matrix called Google matrix. For that matrix, we calculate the Eigenvector that's called the page rank vector.

The dictionary we have created to know page links. The hyperlink matrix H & A matrix created & indexed of all the links we have taken, again Eigenvectors are converted into a matrix. The page rank determines the important web page. Fig. 7 using the PageRank algorithm we can determine the number & of links which is of good quality to a web page to determine the importance of the website or web page. A comparative Study of the pretrained CNN method and existing methods are shown in Table 14. The comparison is done to show that pretrained models are best performer than existing traditional CNN technique.

4.3 Results comparisons of proposed EERF method and pre-trained convolution neural network (CNN) Resnet50

The overall results shows proposed entropy based random forest for image classification is better than CNN and less performer than pre-trained model Resent 50 models. The ensemble based random forest for text classification is better than [21] model giving 99.55% of accuracy.

Table 14. A Comparative study of CNN method and existing methods

Author	Methodology	Accuracy (%)
Li Deng et al., [19]	combination of multiple classifiers	95.7
	Pre-trained Convolution Neural Network (CNN) Resnet50	99%

Table 15. A comparisons of the proposed EERF method and proposed pre-trained convolution neural network (CNN) Resnet50

Classifier	Type of input	Accuracy Achieved
EERF classifier	Image	87
EERF classifier	Text	99.55
CNN	Image	85%
Pre-trained Model Resnet50	Image	99%

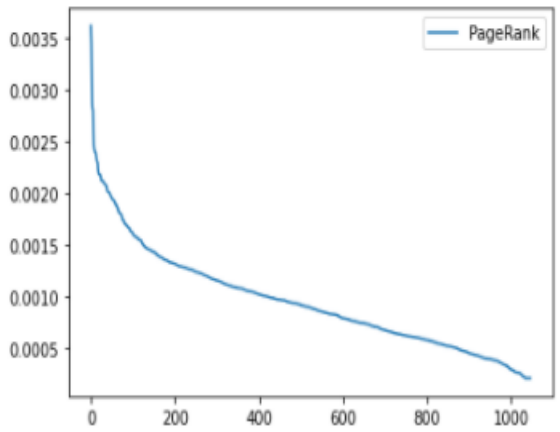

Figure. 7 Web pages with the calculation of the page rank Source: brandeis.edu max difference: (0.0004003690927218846 7+0j) Target: pace.edu

5 Conclusions

In this work, a robust classifier called EERF was proposed for both images and text classification models. The web scraping techniques were used in this study to fetch structured data for three different classes animals, flowers, and food. The image and text data were pre-processed using the natural language processing techniques namely tokenization, and normalization, and it was also used for the extraction process. The important features were selected using color histogram and label encoder techniques. The proposed EERF classifier was used for the final classification of text and image data.

The overall conclusion of the work is the experiments were conducted on three classes to validate the effectiveness of the proposed EERF classifier with traditional classifiers such as SVM, KNN and CNN. The obtained results showed that the proposed EERF classifier for text achieved 99.55% of accuracy. In addition, the proposed method achieved only a 0.6% error rate. The proposed EERF classifier improved the performance in classifying the web pages than existing techniques. The entropy-based random forest for the image classification provides the best performance when compared to the CNN and less performance compared to the pre-trained model Resnet50. The entropy-based random forest for the text classification provides the best performance than all existing models. Future work, an effective feature selection technique will be developed to validate the massive amount of web content from multiple classes of various websites. Some more techniques can be included for improving the recommendations.

Conflicts of interest

The authors declare that there is no conflict of interest

Author contribution

Chaithra, Lingaraju Gowdru Malleshappa and Jagannatha Sreenivasaiah contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

References

[1] N. Aggrawal, A. Ahluwalia, P. Khurana, and A. Arora, "Brand analysis framework for online marketing: ranking web pages and analyzing the popularity of brands on social media", *Social Network Analysis and Mining*, Vol. 7, No. 1, p. 21, 2017.

[2] V Raju, and N. Srinivasan, "Prediction of user future requests utilizing the combination of both ANN and FCM in web page recommendation", *Journal of Intelligent Systems*, Vol. 29, No.1, pp. 583-595, 2020.

[3] A. Prasanth, "Intelligent Recommendation System using Semantic information for Web Information Retrieval", *Advances in Computational Sciences and Technology*, Vol. 10, No. 8, pp. 2367-2380, 2017.

[4] W. T. Chu and Y. L. Tsai, "A hybrid recommendation system considering visual information for predicting favorite restaurants", *World Wide Web*, Vol. 20, No. 6, pp. 1313-1331, 2017.

[5] R. Katarya and O. P. Verma, "An effective web page recommender system with fuzzy c-mean clustering", *Multimedia Tools and Applications*, Vol. 76, No. 20, pp. 21481-21496, 2017.

[6] H. Singh and P. Kaur, "An Effective Clustering-Based Web Page Recommendation Framework for E-Commerce Websites", *SN Computer Science*, Vol. 2, No. 4, pp. 1-20, 2021.

[7] R. Rooba, "Webpage Recommendation System Based on the Social Media Semantic Details of the Website", *Turkish Journal of Computer and Mathematics Education*, Vol. 12, No. 6, pp. 237-243, 2021.

[8] S. Zhang, S. Zhang, N. Y. Yen and G. Zhu, "The recommendation system of micro-blog topics is based on user clustering", *Mobile Networks and Applications*, Vol. 22, No. 2, pp. 228-239, 2017.

[9] R. Wagh and J. Patil, "A Novel Web Page Recommender System for Anonymous Users Based on Clustering of Web Pages", *Asian Journal For Convergence In Technology*, ISSN-2350-1146, 2019.

[10] J. Jiang and H. H. Wang, "Application intelligent search and recommendation system based on speech recognition technology", *International Journal of Speech Technology*, Vol. 24, No. 1, pp. 23-30, 2021.

[11] A. R. Sulthana and S. Ramasamy, "Ontology and context-based recommendation system using neuro-fuzzy classification", *Computers & Electrical Engineering*, Vol. 74, pp. 498-510, 2019.

[12] L. Rajani and U. Thakar, "Webpage Recommendation for Organization Users via Collaborative Page Weight", *Journal of Scientific Research*, Vol. 65, No. 1, 2021.

[13] C. Nigam and A. K. Sharma, "Experimental performance analysis of web recommendation model in web usage mining using KNN page ranking classification approach", *Materials Today: Proceedings*, 2020.

[14] G. Deepak, A. Ahmed and B. Skanda, "An intelligent inventive system for personalized webpage recommendation based on ontology semantics", *International Journal of Intelligent Systems Technologies and Applications*, Vol. 18, Nos. 1-2, pp. 115-132, 2019.

[15] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques", *Expert Systems with Applications*, Vol. 92, pp. 507-520,2018.

[16] D. K. Vishwakarma, D. Varshney, and A. Yadav "Detection and veracity analysis of fake news via scrapping and authenticating the web search", *Cognitive Systems Research*, Vol. 58, pp. 217-229, 2019.

[17] F. Ali, P. Khan, K. Riaz, D. Kwak, T. Abuhmed, D. Park, and K. S. Kwak, "A fuzzy ontology and SVM–based Web content classification system", *IEEE Access*, Vol. 5, pp. 25781-25797, 2017.

[18] S. Wang, L. Sun, W. Fan, J. Sun, and S Naoi "An automated CNN recommendation system for image classification tasks", In: *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 283-288, 2017.

[19] L. Deng, X. Du, and J. Shen, "Web page classification is based on heterogeneous features and a combination of multiple classifiers", *Frontiers of Information Technology & Electronic Engineering*, Vol. 21, No. 7, pp. 995-1004, 2020.

[20] B. Xu, Y. Ye and L Nie, "An Improved Random Forest Classifier for Text Categorization", In:

Proc. of IEEE International Conference on Information and Automation, pp. 795-800, 2012.
[21] A. Gupta and R. Bhatia, "Ensemble approach for web page classification", *Multimedia Tools and Applications*, Vol. 80, No. 16, pp. 25219-25240, 2021.