Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
DVD Part Number: CFP21OAB-DVD; ISBN: 978-1-7281-8381-7

Eliminating the Web Noise by Text Categorization and Optimization Algorithm

S. Ganeshmoorthy
Ph.D. Research scholar, Department of Computer Science
Sree Narayana Guru College
K.G.Chavadi, Coimbatore - 641 105, Tamil Nadu, India.

Dr. R. Priya
Associate Professor & Head, Department of Computer Science
Sree Narayana Guru College
K.G.Chavadi, Coimbatore - 641 105, Tamil Nadu, India.

***Abstract*—In this research, it had been proposed a novel hybrid framework Optimized K-NearestNeighbor (KNN) and FireFly (FF) algorithms to eliminate local noises from web pages. Firstly, the Optimized-KNN is used to classify the contents in the web pages, and second, the FF is used for optimization and noise reduction. The performance metrics for proposed KNN-FF such as Text categorization, Noise classification, and Accuracy show better results while comparing it with the existing NWDL-STC and DOM-SST-ANN techniques.**

Keywords— Noise Removal; Categorization; KNN; Firefly;

I. INTRODUCTION

The WWW has become a common location of knowledge distribution because of the quick growth of the Internet. The material is informative and beneficial. Which consists of over 70 million internet sites in the WWW, effect from the date. WebMining (WM) is used to automatically identify and retrieve trends or knowledge from WWW's records and resources through datamining algorithms. The WM consists of multiple target Web pages [6] to gather data, particularly for Website framework mining and information mining. Here also have to point out that Online customers play a vital role in the process of extracting data or seeking information on the Internet because the WWW is an effectively coordinated tool.

The WWW offers essential knowledge or origin of details for certain web mine activities in terms of the internal contents of web sites. But obviously, this valuable data in web sites also arrives with a variety of noises, such as publicity banners, browsing bars, and reminders of disclosure/copyright. Even though noise data items/blocks helpful for web users and important for site owners as well, the noise often affects automatically extract or collected information, then the WM activities such as the acquisition and extracting information while clustering of webpages and classification of webpages. So, typically, the noise from websites relates to being repetitive, inconsistent, or hazardous to key content. The cleaning of web pages is the preparatory phase and the significant task for webpages to manage this noise material.

Automated attribution of the predefined group of subjects to documents is the categorization of text. The need for categorized text is to enable the valuable retrieval of information by sorting out the unnecessarily complicated things in the webpages, also handling WWW is growing significantly as the number of online texts including webpages grows [7]. This job is usually done by field specialists manually. However, it is doubtful that the human classification can keep up with the pace of development of the WWW. Thus the value of automated categorization of Webpages becomes evident as the WWW continues to grow. In comparison, automated categorization is considerably cheaper and quicker than manual categorization.

The problem formulation of this research discusses three concerns about the traditional KNN categorization used in the classification of webpages: selecting the features, weighting terminology, and similarities between documents and documents.

Initially, a training document has been submitted to traditional KNN classifiers, in each term occurs after elimination of stopwords, but in this optimized proposed KNN classifier only terms of the training document for about the categories that are applied to the document by selecting only optimal features. A Webpage has a clear connection to website material, so the web site material consisting of words and HTML tags that are annotated with terms should be displayed on the web page. Next in this optimized KNN classifier also suggest a scheme of weighing the words with terms and conventional word figures annotated with HTML tags. As the classification of KNN decides the suitable divisions for a new document utilizing k-nearest samples, the resemblance measurement of the document-document may have a decisive impact on the usefulness of the KNN classification. Finally, this optimized KNN suggests a similarity test which was an enhanced version from the traditional KNN, here the similarity of a standard for vector space to pick neighbors that are adjacent to it is the more important training process.

Then the Firefly optimization technique was implemented for eliminating out the local noises from webpages and isolate the key material by utilizing different strategies after classifying the webpages. Thus this developed framework will eliminate local noises from the webpages and produce the required text for users. This research will demonstrate better outcomes while comparing it with existing methods.

The rest of the paper was organized as follows: Section 2 gives about the related works in eliminating the noises from recent researches, Section 3 deals about the methodologies of

the existing and proposed systems for filtering the noises in the webpages, Section 4 discuss the comparison of the results and discussion and Section 5 finally concludes this research article.

II. RELATED WORK

In [1], the researchers stated that the rise of social media comprises a large number of increasingly expanding details. A web-based information retrieval framework and a classification of online data based on a web-based classification are required for the usage of this data. The classification of web pages has many uses, including web directory building and the creation of oriented crawlers. Throughout their research, they discuss the features of webpage grading, they generate an analysis of the literature through the synthesis and evaluation of all webpage classification references, they reviewed different classifiers were using to classify webpages. Eventually, they pursue the principles underneath the techniques reviewed.

In [2] the authors claimed although, in recent days, the web has become the most effective digital tool for the human race because of the development of internet growth. With the Online world rising exponentially, it is the ultimate task to derive valuable material through this web. Meanwhile, the pages identified have various dubious information sections which are not helpful for the consumer and often weaken the method of extracting information. These non-interested structures are commercials, posters, trademarks, window frames, etc which are usually referred to as website noise. The primary role of pre-processing is eliminating such noise in the websites. To retrieve valuable knowledge from a Web a process is needed to removes the noises and almost duplicates them. There are 3 phases in their suggested process. At first, the site page has multiple blocks. In the tag review and DocumentObjectModel (DOM) Tree, a block that is known as noise is omitted. Besides, the computation signatures utilizing revised Simhash techniques with proximity calculation are responsible for the removal of duplicate sections. So many criteria are derived from the various blocks, namely Titleword, Linkword, and Contentword. Hence a graded sequence rating method measures the outcomes for each section by obtaining the important contents. The higher score blocks are collected and the key material is eventually retrieved throughout the website page. The research was conducted and the findings indicate that the suggested system successfully reduces noise.

In [3], the researchers stated that it is incredibly difficult to acquire useful details on the Web while query searching due to the massive extension, multiple complexes, and poor reliability in WWW. Several WM techniques are employed to overcome this problem. The main difficulty is to break rid of the graphic, images, sound, videos, hypertext, etc., that aren't linked to a customer request consists of noisy content or unnecessary details through the website. A modern customized searching framework with an effective algorithm is introduced in this paper to resolve such problems. The proposed PatternExtractor (PE) algorithm based on UniformResourceLocator (URL) (UPE) extracts all related index pages from the site and rates them on a user-based query format. Then the algorithm NoisyDataCleaner (NDC) is used to delete inappropriate material from the pages they have retrieved. The findings illustrate the strong accuracy and retrieval rate of the proposed UPE and NDC algorithm in contrast to current algorithms for the various datasets.

In [4] the researchers stated that many sections of data are found on business websites. In addition to crucial key sections, subordinate sections such as personal data alerts, promotes, patents and links are non-eligible. These sections are referred to as modules of noisy data. The value of Retrieval of Information will decrease the information provided in the noise block. It is a tremendous job to remove such noises. Their work is intended to retrieve the critical details in the websites by noise reduction. After elimination, the material would be filtered and shown in a regular layout. This requires utilizing the web-blocking strategy in which the web-block is separated into ambiguous graphical sections. The graphical sections are categorized according to their characteristics by mathematical similarity measures. The content of the search needed is extracted throughout the website and a clean website is given to the customer. Their solution is to eliminate CrucialNoises, AuxiliaryContents, and Block-based noise information. Per block's value is determined with the Hybrid hash technique. The necessary sections are chosen using the improved Sketching technique concerning the threshold value, which allows it easier for the website to retrieve data efficiently.

The most new progress have included a significant re-weighting methodology [13], the noisy based deep neural-network learning environment[14], and learning for image classification from large noisy results [15], a robust neural network with cross-entropy loss [16], correction of losses [17], among several others. Losses or specimen rectification also has been analyzed in the literature of learning with weak supervisors with unlabeled data [18]. Almost all of these operate whether in a total absence of theoretical assurances against asymmetric noise rates from the proposed method [19] or necessitate noise rate estimation. The latest study [20] suggests a theoretical loss of information, an idea adapted from an earlier theoretical contribution that is somewhat reliable to asymmetric noise rates. Researchers [21] targeted for such a simple loss function to be optimized that could quickly integrate to current ERM solutions.

III. METHODOLOGIES

a. EXISTING MODELS:

(i) Noise Web Data Learning (NWDL) and Suffix Tree Clustering (STC)

A data mining technique is given in this segment which can learn noise in a website's UserProfile (UP) until it is removed. The main emphasis is the learning, detection, and removal of noise, taking account of an individual's diverse interests and the changing site data [5]. Noise reduction is

focused on what is and is not interesting in a user in the extracted site log info. The interest of a customer in a web page is often addressed in a recent analysis, due to the amount of time that they visit this web, the amount of time they interact on the internet, the number of connections, and the latest visits. To a certain degree, recent studies evaluate user interest in web data logs extracted however there is little evidence to prove how noise is observed in a web UP before removal. Fig. 1 describes this work.

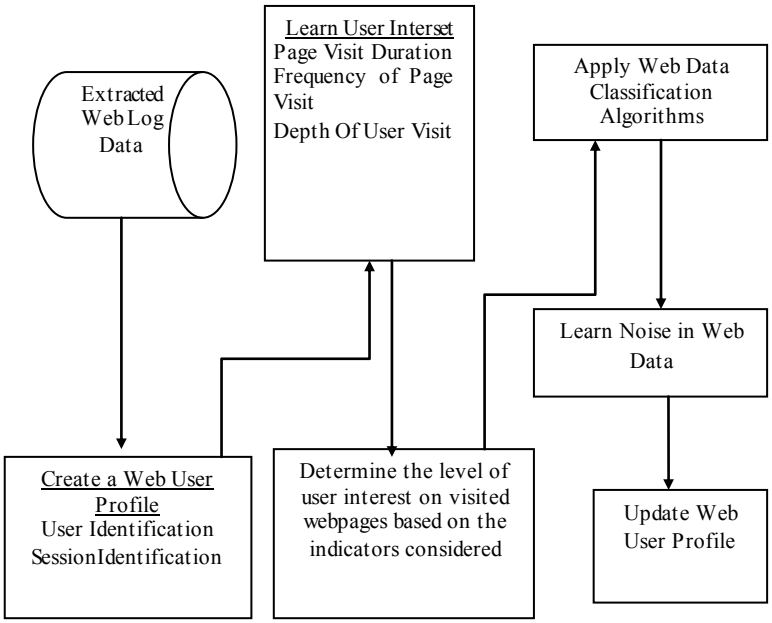

Fig. 1. NWDL Framework

A UP seems to have a group of URLs that are of concern to the consumer. The development of a UP is focused on several webpages that are visited by a user concerning its importance. A selection of sessions is used on UP. This work would discover the degree of consumer engagement on the pages visited and assess valuable knowledge from noise data after building the UP. Different steps, i.e. duration, length, and scope of user visits to a website are recognized.

The STC is a linear time cluster algorithm focused on the phrases common to text classes. A structured series of one or more words is a term in our context [8]. Here it identified a base cluster as a series of documents with a common sentence. STC is divided into three logic steps:

- Cleaning of a document
- BaseCluster recognition using SuffixTree and
- Clustering

By these stages, each document's text string is modified to a LightStemming technique (eliminating the suffix and prefix of each word then minimizing Pluraly to Singularly). The limits of the sentence (recognized through HTMLTags and punctuations) are labeled and tokens that are NonWord (such as the digits, the HTML tags, and much of punctuation) are stripped. BaseCluster recognition may be interpreted as the development of an inverted phrase index for that text set. First, the proposal was presented to use a SuffixTree to cluster records. In this case, it has an enhanced clustering algorithm, which implements the merging of simple clusters (stage 3 of the STC) and compares them to the classical clustering approaches in the Web domain using a common technique. This was developed as the first phase [11] for web noise removal research work.

(ii) Document Object Model (DOM) with Site Style Tree (SST) and Neural Network (NN)

The DOM framework is a platform that the WWW Consortium (W3C) produces as a tree structure in the memory as an HTML and XML code. Any software allows users to access these XML documents in its memory tree which represents a copy of the layout of the data. The DOM enables the consumers to modify the XML document dynamically and edit it. It gives the entire document as a model, not just an HTML tag [9]. It provides the tree structure for the documents. These trees can effectively be used with a full webpage and are extremely able to adapt. These trees are the HTML document model that is well defined. However, these trees are adequate to reflect a single HTML page's layout/PresentationStyle (PS), the general PS and content of a group of HTML pages cannot be checked and cleared based on specific DOM trees. DOM trees in this research work are also not enough to take into consideration both the appearance of the PS and their actual content.

Thus in this research the noises are termed as follows: (i) the further PS does an ElementNode have is not much significant (ii) Almost much necessary content having by ElementNode is much significant. In determining the significance of an ElementNode, these two essential values are used. The presentation aims to detect noises with frequent PS, while the value of content is to define the key content of the pages which might be displayed in such types. Thus the significance of an ElementNode is provided in this method by combining the importance of its presentation with the importance of its contents [10]. The more important a node is the most often the key material would be the most important in it.

Unwanted data units typically share several web sites of the same material and presentation style. A framework called SiteStyleTree (SST) was designed to collect this at the site level. When the SST for a webpage is constructed, it is simple to distinguish between informational and non-informational elements.

The shades of the SST are more prone to noise in the illustration shown in Fig. 2 as the PS is extremely normal, set, and thus less significant (alongside the relevant information not given in the figure). There are several child-style nodes in the double-line table ElementNode, suggesting that the ElementNode is possibly significant. That is, it is more probable that the double-line table would include the key pages. In particular, the text ElementNode with two lines is also important, because its content is varying, but its PS is constant. Enable the SST to be the ST generated from all website pages.

The NN architecture was combined with this DOM and SST to enhance noise reduction. The NN may be recognized

for the usage of traditional algorithms as a successful task solution process. When its source is something that had been never used before, it creates a result that is identical to that of the nearest similar pattern of input training. Each input layer node gets source features, operates, and transfers towards the following layer in the NN architecture [9]. The weight that is the intensity of relation among 2 nodes is used for this method. A NN is a system consisting of a set of basic nodes or elements identified as neurons. These components still function concurrently. The NN's role is primarily defined by the neuronal link. The neurons are connected, and the values referred to as weights change in each connection. The weight update method is called learning. The NN classification outcome is used to remove multiple noise patterns.

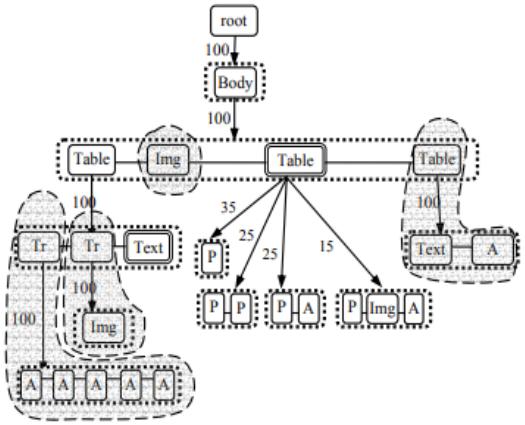

Fig. 2. Structure of SST

This was developed as a second phase [12] for web noise removal research work. While comparing the second phase with the first phase the second phase proves better results, to enhance the efficiency of the research output the third phase was proposed with meta-heuristic algorithms which was able to process the outcome with better time efficiency for a large number of documents as well as with improved accuracy.

(b) **PROPOSED MODEL**

In this proposed model the Optimized K-NearestNeighbour (KNN) classifier was used for the classification of the webpages and the optimization algorithm FireFly (FF) was used for the noise removal process. Fig. 3 shows the workflow for the proposed model.

(i) Optimized KNN for Classification:

Almost every training document is typical in traditional KNN classifier as the terms of the document after the stopwords have been deleted, although it cannot be stated that all the terms are linked to the training document's categories. Noise in the text categorization task may be the circumstances and are not aligned with the groups of the training manual. Therefore unlike traditional KNN classifiers, the optimized KNN classifier proposed in this research decreases noise levels by utilizing selecting an optimal features process. It initially chooses the index terms (features) that forecast the occurrence of that category for any category to eliminate the noise of a training document and delete all words not chosen as features from the categories of the training material allocated. The optimal result of this proposed process of selecting the features is to exclude all noise words that do not apply to groups under which the index terms appear in the documents under the training process. Since the collection of features decreases the dimension of the vectors, the closest training document can be located in the proposed optimized KNN classifier quicker. Also, training documents of the same categories with related term vectors are defined to locate them tightly within a vector field. This selection of features result suggests a good impact on categorizing the content.

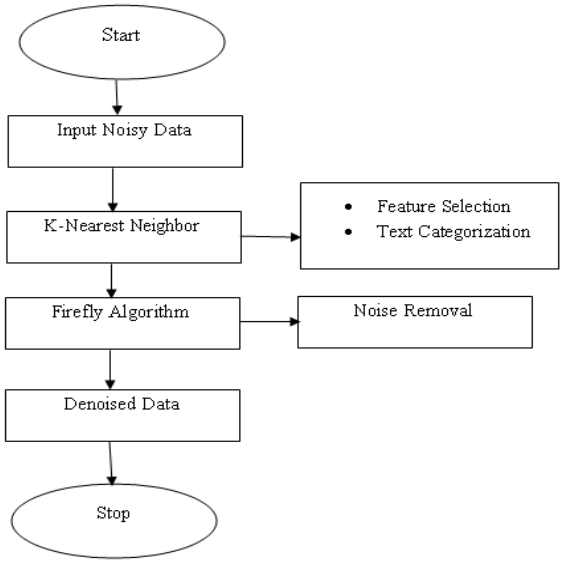

Fig. 3. Work Flow for Proposed Model

Here it selects measurements for Expected Mutual Information (EMI) and Mutual Information (MI), which are typically used for categorizing documents, template identification, and artificial intelligence, for the selecting features process. If the EMI and MI values between the word and category are above the user-specified threshold of EMI and MI respectively, it shall pick an index term as a function for a category. EMI is obtained using the similarity and distinction, as defined in Equation (1), between an index and a group.

$$EMI(Ti,Ck) = \sum_{a=0,1}\sum_{b=0,1} \mathrm{P(Ti = a, Cj = b)} \left(\log \frac{\mathrm{P(Ti = a, Cj = b)}}{P(Ti = a) * P(Cj = b)}\right) \quad (1)$$

The co-occurrence window in the following Equation (2) and (3) is fixed inside a document. Under Equation (3), apart from features with positive characteristics but even negative characteristics may be obtained for each type.

$$Weight(Tagk) = UDW(Tagk) * DR\ (Tagk, Di) \quad (2)$$

Here the weight defined by the user tag UDW(Tagk) was based on the "Tagk" Mark-Up tag.

$$DR(Tagk, Di) = \{\frac{1,\ if\ \beta \le \lambda}{\lambda/\beta, otherwise} \quad (3)$$

$$\text{where } \beta = \frac{\text{the frequency of Tagk in document Di}}{\text{the total frequency of all tags in document Di}} \quad (4)$$

$$\lambda = \frac{\text{the frequency of tag Tagk in a collection}}{\text{the total frequency of all tags in a collection}} \quad (5)$$

Equation (6), pick the tag containing the highest weight if a term is indicated with many tags. Also while describing Eq (6) if a term happens in a mark with a weight, the term frequency is numbered a period rather than '1', and a logarithm formula is used to compute the term weight as follows:

$$W_{TpDi} = \frac{\log\left(\frac{N}{dfTp}\right)}{\log(N)} * \frac{\log\left(\sum_{j=1 Tagx \in TAGTp,j}^{tfTp,Di} \max(\{Tagx\})\right) + 0.5)}{\log\left(max_{Tq \in Di}\{\sum_{j=1 Tagx \in TAGTq,j}^{tfTOq,Di} \max(\text{Weight (Tagx)})\}\right) + 1.0)} \quad (6)$$

Here,

- For the "Di" document based on the "Tp" term the frequency of the term is "tf_{TpDi}".
- In the "Di" document, for the "Tp" term the occurrence is "jth" which was annotated by Mark-up Tag sets "$TAGT_{p,j}$".
- For a collection, the document quantity is "df_{Tp}".

Since features of negatives in the KNN classification are not beneficial, it prefers features only with positive characteristics with following Equation (7) for defined MI:

$$\text{MI}(Ti, Ck) = log\frac{P(Ti = 1, Cj = 1)}{P(Ti = 1) * P(Cj = 1)} \quad (7)$$

For EMI the term is chosen as a good attribute for the group if the correlation between a term and a group is much greater than the differences. The variations between the word and the group suggested that it is more possible for other groups to pick the expression as a good attribute. The features of positives extracted by EMI thus appear to estimate fewer groups than the feature of positive extracted by MI. The further categories the more the word predicts. Therefore the features with positive attributes of each type were not only gained by MI.

(ii) FireFly for Noise Removal:

Training with noise means that a weight dependent on an approximate likelihood of class noise is applied to every other training set. When training from the original noisy training results, the FF model would assume the class noise rate. Therefore, this technique aims to identify and delete inaccurate data from the training collection. The rate of class noise in the training data can be minimized by this process. Features are individual terms or properties in this feature collection for optimization (words, image files, link references, etc). Things in Web pages, though, have such constructs that are represented by its embedded HTML tags.

This method is influenced by the light of the FireFlies (FF) and shows the relationships between the FF. There are around 2000 species of FF, the majority of which emit rhythmic brief flashes. For a specific FF, the sequence of flashes is also special.

A summarization and a brief description of the pseudocode is given. The several webpages or websites are taken as input into the FF method. The webpages are accessed one at a time. When the webpages have been read, the HTML tags are reviewed in phase 3 and the document with many tags would be considered. Phase 5 measures the target function and generates the initial FF population in phase 6. In phase 7, the LightIntensity is defined and in step 8, the absorption coefficient is defined. The maximum performance would be calculated in phase 9 based on the latest approach to update the LightIntensity. The information with noise is remembered and removed in phase 18 and phase 19. The key contents are eventually extracted.

Both data on webpages are processed using the FF technique to effectively recover the pattern. The database is designed utilizing a network to store the corresponding data on websites. Complying the installed DOM tree with the database separate from noisy data with exclusion. Finally, the key material may be accessed. Initialize the objective feature "f(wi)", which differs according to the inverted square regulation, by the LightIntensity "I(o)". The following Equation (8) is used for this process:

$$\text{I(o)} = \frac{Is}{Io} \quad (8)$$

The intensity of "I(o)" is at the source and 'r' is the distance of the observer. The LightIntensity 'I' differs by the distance square 'd'. The coefficient of absorption 'μ' is determined by the given Equation (9):

$$I = I_o e^{-\gamma a^2} \quad (9)$$

For removal of the noises from the webpages using FF was given below:

Given Input: Several webpages from sites.

Expected Output: Only to produce exact information from the webpages without noisy contents.

Phase 1: Provision for accessing several webpages

Phase 2: Ability to analyze each page

Phase 3: Confirming the tags of the HTML

Phase 4: Able to considering multiple tags for a given document

Phase 5: Setting f(wi) w=(w1,w2,w3..) as an Objective function

Phase 6: Creating the FF population preliminary

Phase 7: LightIntensity get formulated
Phase 8: 'γ' Coefficient for absorption get defined
Phase 9: When (t < Max_Generation)
Phase 10: For i=1:n
Phase 11: For j = 1:n (n fireflies)
Phase 12: If(Ij> Ii)
Phase 13: The 'i' FF gets moved to 'j' FF
Phase 14: The LightIntensity gets updated and new solutions get calculated
Phase 15: End if
Phase 16: End for 'j'
Phase 17: End for 'i'
Phase 18: Information with the noise was identified
Phase 19: Noise gets Eliminated finally

FF's attractiveness is equal to that of another FF's expected LightIntensity. Brightness noted by an adjacent FF 'β' with the formula calculated:

$$\beta = \beta_o e^{-\gamma d^2} \tag{10}$$

Then, launch the population of the FF. FF 'i' draw the most attracted FF 'j'. The movement is measured by the given equations:

$$x_i = x_i + \beta\left(x_j - x_i\right) + \alpha\varepsilon \tag{11}$$

In the accompanying formulation, the fitness method for extracting noise from websites is calculated as:

$$\text{Fitness} = \alpha\frac{T_{tot} - T_{neg}}{T_{tot}} + \frac{\beta}{F} \tag{12}$$

Here the "Ttot" is presented as a cumulative amount of tags on a webpage, "Tneg" is a Negative-tag on the webpage. The FF attractiveness was termed as 'β' and F-measure was termed as 'F'.

IV. RESULTS AND DISCUSSION

The approach outlined in this paper has been implemented as a Firefox toolbar, including all the algorithms mentioned. Firefox has been chosen since it is amongst the quite popular and commonly used browsers, even though it is open source and unrestricted. The XUL is an XML-based language used to enforce the interface, and Javascript, which implements the toolbar activity and behaviors, is being used to implement as a Firefox toolbar. It comprises 2577 LOC in particular. The consumer will access the Internet as usual with this app. Thus, users just need to click a button anytime they want to remove the template from a website and the tool automatically loads the correctly connected webpages to form a CS. In the browser, the connections to all web sites are then shown.

This approach can be applied by inputting a website (called the main page) in real-time and outputting a series of web pages incorporating (a portion of) the same design. This establishes a full subdigraph in the topology of the website to explore these webpages. A subdigraph describes a list of web pages that are connected pairwise. For eg, the key material, i.e. the news, is within the pagelet in the dashed square of the news webpages. Along the main material, there is a "Popular this week" pagelet with the most important articles and another pagelet for subscriptions and social networks. Consequently, between the menu and the main content, a selection of similar news (various for each webpage) is shown.

Based on the TextCategorization, Accuracy, and NoiseClassification, the results are obtained. The work proposed was compared with the existing models. Table 1 and Fig. 4 shows the categorization of the text. Here, the efficiency of the website classification is evaluated based on the optimized KNN classification. The training samples are divided into Training-sets (90% of the exact Training-settings) and a Test-set (10% of the exact web's HomePages) this was done for validating the optimized KNN approach by optimizing the 'K' and the thresholds of EMI and MI for selecting the optimal features. Here it used an OriginalCosine SimilarityMeasure to find KNN for a webpage while research validation is carried out. The percentage in the table denotes the categorization rate for a given dataset.

TABLE 1: TEXT CATEGORIZATION

S.No	Methods	Text Categorization (%)
1	NWDL+STC	78
2	DOM+SST+ANN	89
3	KNN+FF	94

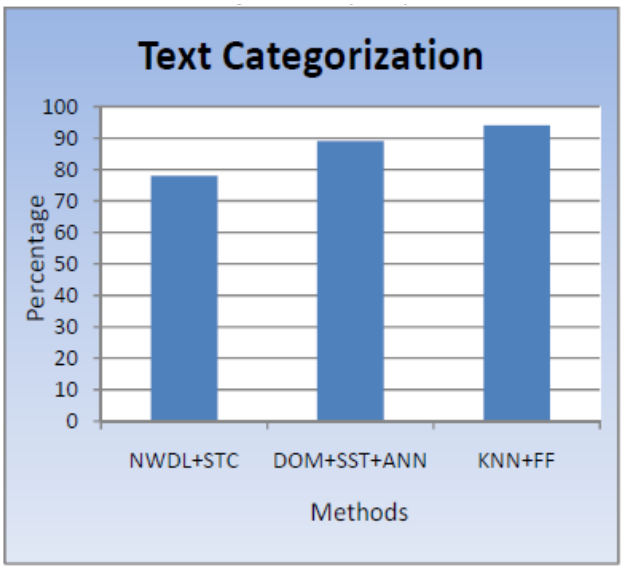

Fig. 4. Text Categorization

The noise rate is based on the true label of a sample in this model and is independent of the sample's feature set. In most other terms, various labels are prone to flip to the wrong label with different odds. Provided that x is the collection of

features of a sample, letting y~ be the observed x label, y is the true sample label of x and p is the likelihood of flipping the true label into such a noise label (thus p is the class noise rate or noise rate for short).

Comparing to existing methods, noise reduction from the website in UPs is calculated based on the shift in consumer interest over time and noise trends. However, when time prolongs, the category has been visited by a person, where current noise data trends are used to remove noise in the webpage. A viewing a webpage alone cannot assess the duration and frequency of a visit. A consumer can be only involved in a specific amount of time. Consequently, visit frequency and length are not obvious when noise is excluded, which was shown in Table 2 and Fig. 5. The value is obtained by the software by inputting the webpages.

TABLE 2: NOISE CLASSIFICATION RATIO

S.No	Methods	Interest	Noise
1	NWDL+STC	146	82
2	DOM+SST+ANN	146	96
3	KNN+FF	146	98

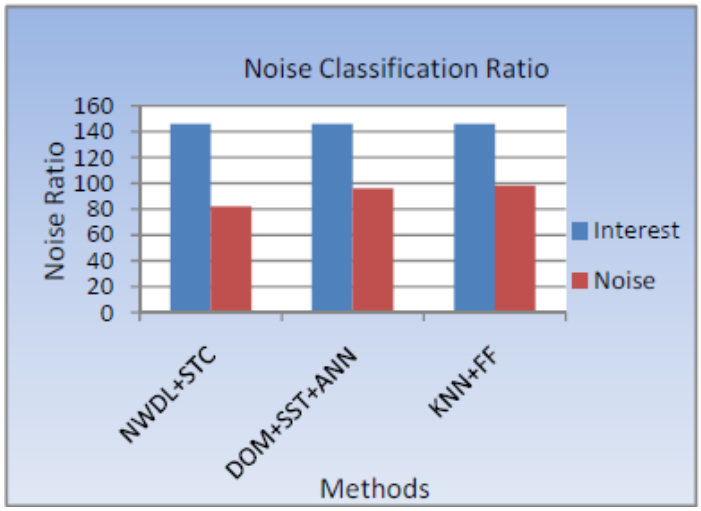

Fig.5. Noise Classification

The classifier's accuracy is based on determining the class of a new example is understood as efficiency, while noise robustness was being described as the accuracy loss rate of the classifier caused by the existence of noise density, in the case of non-noise. Through 5 runs of a stratified 5-fold cross-validation, the formulation of accuracy calculation of each methodology is obtained. The data set (web pages) is split into five partition sets with equivalent instance numbers and the proportion between groups is retained in each fold. As a comparison for the classifier learned from the four remaining partitions, each partition set is used. The performance of the accuracy for the noise reduction method relies not only on the accurate classification outcomes of the FF but also on the categorization by the optimized KNN in depth. Multiple structures were divided into heuristic structures since webpages for numerous websites are designed with a complicated and different framework. The accuracy rate is 99 percent while the KNN-FF algorithm is combined. The value of the accuracy was obtained by the tool for different webpages. Table 3 and Fig. 6 shows the accuracy rate of this research work.

TABLE 3: ACCURACY

S.No	Methods	Accuracy (%)
1	NWDL+STC	90
2	DOM+SST+ANN	97
3	KNN+FF	99

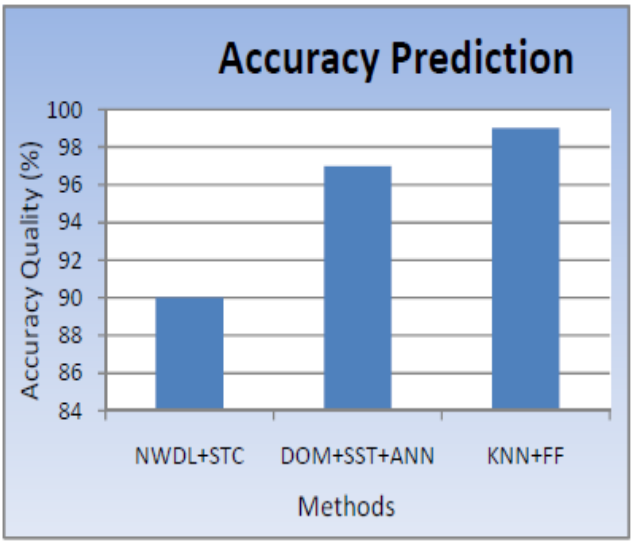

Fig. 6. Accuracy

V. CONCLUSION

In this research, it had been proposed a novel hybrid framework Optimized K-NearestNeighbor (KNN) and FireFly (FF) algorithms to eliminate local noises from web pages. Firstly, the Optimized-KNN is used to classify the contents in the web pages, and second, the FF is used for optimization and noise reduction. The Optimized KNN will enhance classification efficiency by using advanced similarity metrics. Then the Firefly optimization technique was implemented for eliminating out the local noises from webpages and isolate the key material by utilizing different strategies after classifying the webpages. Thus this developed framework will eliminate local noises from the webpages and produce the required text for users. This research demonstrated better outcomes while comparing it with existing methods. Noise classification and Accuracy of the proposed optimized KNN-FF show better results while comparing it with the existing NWDL-STC and DOM-SST-ANN techniques. The limitation is thus, analyzing their success in the experiments in-depth, we could see that it was responsive to algorithm parameters like the k value in the kNN and the perception training rate as well as the algorithm's

termination point, it couldn't even operate on the clean data. We may discuss noise handling activities such as semi-supervised grouping, transductive transfer learning, and also examine the issue of domain adaptation in future works. For various groups, we will often include different noise rates, as label noise rates, in reality, are also class label based.

References

[1] S.Lassri, H.Benlahmar, A.Tragha (2019), "Machine Learning for Web Page Classification: A Survey", International Journal of Information Science & Technology – iJIST, ISSN : 2550-5114 Vol. 3.

[2] A. Saravanan, S. Sathya Bama (2020), "Extraction of Core Web Content from Web Pages using Noise Elimination", Journal of Engineering Science and Technology Review 13 (4) (2020) 173 – 187.

[3] Pradeep Sahoo, Rajagopalan Parthasarthy (2018), "An Efficient Web Search Engine for Noisy Free Information Retrieval", The International Arab Journal of Information Technology, Vol. 15, No. 3, May 2018.

[4] Uma, R., Latha, B. Noise elimination from web pages for efficacious information retrieval. Cluster Computing 22, 14583–14602 (2019). https://doi.org/10.1007/s10586-018-2366-x

[5] Julius Onyancha, Valentina Plekhanova (2018), "Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:12, No:1, 2018.

[6] N. Pradhan, M. Gyanchandani and R. Wadhvani (2015), "A Review on Text Similarity Technique used in IR and its Application", International Journal of Computer Applications 120(9) (2015).

[7] J. Onyancha, V. Plekhanova, and D. Nelson, "Noise Web Data Learning from a Web User Profile: Position Paper", in Proceedings of the World Congress on Engineering, 2017, vol. 2.

[8] Luyi Feng, Yin Kia Chiam, Erma RahayubintiMohd Faizal Abdullah, UnaizahHanumObaidellah (2017), "Using Suffix Tree Clustering Method to Support the Planning Phase of Systematic Literature Review", Malaysian Journal of Computer Science. Vol. 30(4), 2017.

[9] Thanda Htwe, Nan Saing Moon Kham (2011), "Extracting Data Region in Web Page by Removing Noise using DOM and Neural Network", 3rd International Conference on Information and Financial Engineering IPEDR vol.12, 2011.

[10] Lan Yi, Bing Liu, Xiaoli Li (2003), "Eliminating Noisy Information in Web Pages for Data Mining", SIGKDD í03, August 24-27, 2003.

[11] S. Ganeshmoorthy, Dr.R.Priya. "Enhancing The Web User Profile's Quality By Noise Web Data Learning (NWDL) And Suffix Tree Clustering (STC)". JCR. 2020; 7(19): 4294-4301. doi:10.31838/JCR.07.19.504

[12] S. Ganeshmoorthy and R. Priya, 2020. "Web User Profile Improvisation by Sampling Site Style Tree With Dom Structure and Neural Network". International Journal of Advanced Research in Engineering and Technology (IJARET). Volume:11,Issue:10,Pages:161-170.

[13] Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence, 38(3):447–461, 2016.

[14] Song, H., Kim, M., and Lee, J.-G. Selfie: Refurbishing unclean samples for robust deep learning. In International Conference on Machine Learning, pp. 5907–5915, 2019.

[15] Yi, K. and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7017–7025, 2019.

[16] Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.

[17] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1944–1952, 2017.

[18] Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. arXiv preprint arXiv:1808.10585, 2018.

[19] Jacob, I. Jeena. Performance Evaluation of Caps-Net Based Multitask Learning Architecture for Text Classification. Journal of Artificial Intelligence 2, no. 01 (2020): 1-10

[20] Xu, Y., Cao, P., Kong, Y., and Wang, Y. L dmi: An information-theoretic noise-robust loss function. NeurIPS, arXiv:1909.03388, 2019.

[21] Kong, Y. and Schoenebeck, G. Water from two rocks: Maximizing the mutual information. In Proceedings of the 2018 ACM Conference on Economics and Computation, pp. 177–194. ACM, 2018.