IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WEB PAGE INFORMATION EXTRACTION TECHNIQUE BASED ON WEIGHTED FREQUENCY OCCURRENCE

[1]Ms. Dipali Balshiram Shete, [2]Dr. Sachin Bojewar

[1]PG Scholar, [2]Professor & DAO, Department of Information Technology VIT, Mumbai

[1]Computer Engineering,

[1]Alamuri Ratnamala Institute of Engineering & Technology (ARMIET), Thane, Maharashtra, India

Abstract: The web pages have rapidly increased their applications in online. The increase of bloghosting and content management systems such as blogger, wordpress and tumblr were contributed to the growth as they made possible for users with no experience in maintaining the digital system to share a various content. Recently, in the domain of multimedia and computer vision, a various machine learning and data mining algorithms for automatic image annotation were developed. However, the limitation is partly explained because multimedia content mining was a highly complex that regularly demanded for in-depth computations and larger training datasets to build an effective model. Further, the set of web page categories that a system must handle depends on the specific application and may vary over time that became a barrier for the use of classification methods. In order to overcome such problem for extraction, browse and manage the large number of web images, the proposed a Weighted Frequency Occurrence (WFO) method which performs probabilistic formulation for image extraction and context extraction together. Initially, URLs were collected from the E-commerce webpage such as Flipkart, Ebay and Myntra. Then, the preprocessing process such as tokenization is carried out to breakdown the complex data. The WFO is calculated by using the frequency of words and maximum number of similar words present in the description. The calculated words are used in Term Frequency Inverse Document Frequency (TF-IDF) to categorize the words and to index the terms present in the sentence. Finally, text summarization of the proposed method is carried out by using Latent Dirichlet Allocation (LDA). The classifiers such as Support Vector Machine (SVM), Random Forest and Gaussian Naïve bayes are used in the proposed method. The proposed method is simple and easy to implement and does not require much training data. The experimental result shows that, Gaussian naïve bayes classifier shows higher performance of 92% in terms of accuracy, precision, recall and F1-Score.

Index Terms **- Web pages, Weighted Frequency Occurrence, Term Frequency Inverse Document Frequency, Text summarization, Latent Dirichlet Allocation**

I. INTRODUCTION

Because of the quick development in Internet Technology (IT), webpage has become a most significant platform for the people to communicate and gain knowledge [1]. The website is playing an important role by providing informations to users. The website is limited by fixed Uniform Resource Locator (URL) and displays the content of site as time-variant [2]. With the growth of internet, twitter, blogs, short messages and wechat were acquired popularity [3]. Further, the users of social media platforms such as facebook, whatsapp and twitter started sharing wide range of data in the form of images, texts, video and audio. In addition, the data of various domains like E-commerce, E-learning, and E-government started growing due to the rush of technological device and user. So, it is required to have an information extraction system [4].

The information extraction is the method of extracting the query specific text, images or multimedia contents from the websites where the time of extraction and accuracy were considered [5]. The web pages are made up of HTML elements and includes data among these elements [6]. The information extraction is the method used to identify the exact part of an HTML document which contains the main informations of a web site [7]. The images play the important role for the user experience on web because they allow users to transfer huge amount of informations at a minimal glance. Also, images will capture the attention of audience than texts [8]. The image extraction is presently an important topic and a key method in computer vision. Many researchers are working to develop an efficient information extraction system on image or multiple media database. It is popularly used in image classification, image management, video surveillance and re-identification of pedestrian. The Text based image extraction technique includes the same objects of image extraction and the same category of image retrieval [9,10]. The main goal of data extraction from the web is to control the heterogeneity and to remove the noise in order to provide fast and accurate information to the user [11].

There are several existing web image extraction methods which provide the semantic information's that contains in the visual features of image. Because it is very difficult, consumes time, intensive to labor, and expensive to manually label the large set of images [12]. Mining multimedia contents was much complex because that demanded intensive computation and extensive training dataset to build effective method. By contrast, in order to overcome such problem for extraction, browse and manage the large number of web images, the proposed a WFO method which performs probabilistic formulation for image extraction and context extraction

together. Initially, URLs were collected from the E-commerce webpage such as Flipkart, Ebay and Myntra. Then, the preprocessing process such as tokenization is carried out to breakdown the complex data. The WFO is calculated by using the frequency of words and maximum number of similar words present in the description. The calculated words are used in TF-IDF to categorize the words and to index the terms present in the sentence. Finally, text summarization of the proposed method is carried out by using LDA.

II. LITERATURE REVIEW

Daniel López-Sánchez et.al.[13] developed a framework for webpage categorization based on the visual contents by using metric learning and deep transfer learning. The developed method initially, from the URL extracted every image present in that particular web page and filtered the images which does not contained any selective information. Then, extracted a feature from each image by using Deep Convolutional Neural Network (DCNN). At last, analyzed each feature extracted from web page and combined them to predict the category of the whole web page. The developed method showed higher classification accuracy. However, the developed method was unable to deal with the web page which includes textual information.

Gerard Deepak et.al.[14] developed an enhanced hybrid semantic algorithm to facilitate ontology model for homonyms and related synonyms terms. By utilizing semantic similarity computation and description logics semantic and also established dynamic path to get the recommendation for web images. The developed method classified ontologies based on the query terms by utilizing hard margin Support Vector Machine (SVM) and lookup directory efficiently located homonyms which eased the image recommendation. API measure was utilized to compute the semantic similarity based on dynamic path. The developed method achieved personalization by providing the click through information about images and prioritizing them during recommendation of images from web. However, the datasets used in the developed method was included many irrelevant images.

Farman Al et.al.[15] developed a fuzzy ontology and SVM based system to systematically classify the web content to find and block the access to pornography. The developed method classified URLs into medical URL and adult URL by blacklisting the censor webpages. Then, the developed fuzzy ontology extracts the web content to know the website type such as normal, medical and adult content to block the pornographic contents. The developed method showed efficient performance by automatically identifying and blocking the pornographic content. However, the developed ontology included many imprecise terms, and it was difficult to treat vague information in the web content.

Changqin Huang et.al.[16] developed Multiple-concept Retrieval by using Bi-modal Deep Learning (MRBDL) for web image retrieval. The developed method boosted the differntiality of the multiple concept scene and utilized the multiple concept Fully-Connected (FC) classifier layers in the visual and text CNN to identify the holistic scenes with individual visual characteristics. The semantic correlations of concepts were used in order to improve the multiple concept scene detection. The log likelihood functions of the relevant score were maximized over the training images to maintain the changing frequency of the method. The developed method increased the retrieval performance of method. However, the textual images were not used in the retrieval process due to higher noise.

Amol P. Bhopale et.al.[17] developed swarm optimization-based cluster framework for the retrieval of information from the world wide web. The developed method included 2 subtasks in the preprocessing step. Initially, decomposed the collected data into groups by using K-flock clustering algorithm. Then, extracted the frequent pattern from every cluster by utilizing recursive elimination algorithm. In the next step, implemented cosine similarity based on probability model to extract query specific information from the clusters based on matching score among similar frequent patterns of query and cluster. The existing methods increased the quality of retrieved documents and improved the retrieval performance. However, the existing methods not included the semantic relatedness among the similar frequent patterns of clusters and query so this affects the final retrieval results. In order to overcome such problem for retrieve, browse and manage the large number of web images, the proposed research performs probabilistic formulation for image retrieval and context retrieval together with theoretical arguments that performs extensive experiments for illustration.

III. PROPOSED METHODOLOGY

This section describes the proposed framework, the global structure and detailed description of the framework are outlined. The proposed system performs the following tasks. Initially, the Uniform Resource Locator (URL) searches the website for extracting, identifying the meaningful content or image from Flip kart, eBay, Myntra. Webpages contain different types of scripts, such as the style sheet, the title, and metadata information. The proposed system extracts this information, and morphological analysis is then employed to identify the various forms of words based on HTML/XML Script Reading. The beautiful soup Application Programming Interface (API) searches engine query to retrieve the data such as product's image, rating, description, name and review from the webpages. After retrieving the data, preprocessing process is carried out by tokenization. Then, weight is calculated from the frequency and maximum number of words. The calculated weights are further employed in TF-IDF to categorize the data and to index the terms present in the sentence. The text summarization is carried out by using LDA for extracting the summary from classifiers such as Support Vector Machine (SVM), random forest, Gaussian naïve bayes. The block diagram of the proposed method is shown in Figure 1.

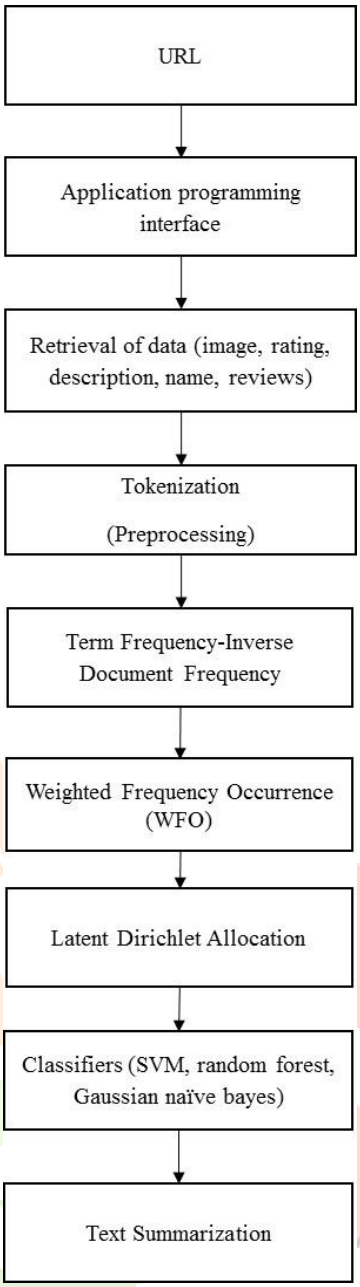

Figure 1. Block diagram of proposed method.

URL

URL is a source of web resources which specifies its own location on the network of computer and method for extracting a location. URL is also known as a web address which contains many parts such as domain name and protocol that explains the web browser where and how to extract the resources. The URL includes the protocol name that need to be accessed a resource and name of the resource. The initial part of URL finds which protocol to use for an access medium. The next part determines the IP address, name of the domain and subdomain from the location of resources. In the proposed method, URL searches the websites such as Flipkart, Ebay, Myntra for extracting the product's image, rating, description, name and reviews present in the webpage. The URL considered in the proposed were,

(1) https://www.flipkart.com/clothing-and-accessories/dresses-and-gown/dress/women-dress/pr?sid=clo,odx,maj,jhy&otracker=categorytree&otracker=nmenu_sub_Women_0_Dresses
(2) https://www.ebay.com/b/TV-Video-Home-Audio-Electronics/32852/bn_1648392
(3) https://www.myntra.com/women-kurtas-kurtis-suits

Morphological analysis

After extracting the contents and images from the webpage, morphological analysis is employed to identify the various forms of words based on HTML/XML Script Reading. Morphological analysis is a technique used to analyze the structure and part of words like root words, stem, suffix and prefix. It also considers the intonation; parts of speech and the way contents change the meaning and pronunciation. The morphological analysis is characterized by 3 steps they are described in the following.

- The matrix and solution spaces were created.
- Identified the possible configuration and consistency constraint.
- Evaluated and selected the most productive solution.

Application programming interface

After employing the morphological analysis to identify the various forms of words, the beautiful soup API searches engine query to easily retrieve the data from webpages. The beautiful soup is library of python which retrieves the data out from HTML and XML files. The API is an interface that describes the communication among various intermediate software. Initially, the search query is sent to API in the form of application to the web server by URL, which include request verb, header and request body. When the API receives search query, it calls to the web server or external program. Then, the server responds to API with requested information. At last, API responds to the initially requested application. If the data is not completely downloaded the URL again checks for the content till the data downloading is completely done. In proposed method, the beautiful soup API retrieves the 22 product's image, description, rating, name and review from the webpage.

Tokenization

After retrieving the data from webpages using beautiful soup API, the preprocessing process such as tokenization is carried out to break the complex data such as paragraphs into small units called as tokens. The tokenization includes 3 terms such bigrams, trigrams and ngrams. where, tokens which includes 2 consecutive words are called as bigrams. Tokens which includes 3 consecutive words are called as trigrams and tokens which includes N-number of consecutive words are called as ngrams. For the collected words from the tokenization, weights were calculated by using frequency and maximum number of similar words in the product description. The weighted frequency occurrence is calculated by using equation (1)

$$WFO = \frac{frequency}{maximum\ number\ of\ similar\ words} \tag{1}$$

Term Frequency-Inverse Document Frequency(TF-IDF)

The calculated weights are employed in TF-IDF to categorize the words and to index the terms present in the sentence. TF-IDF is a technique that categorizes the textual data and indexes the terms as per the requirements of documents. The TF-IDF is based on the frequency and vice versa of the term present in the sentence. The TF-IDF is calculated by using the equation (2)

$$td - idf = tf(t_i, d).idf(ti) \tag{2}$$

Where t_i is ith term in sentence, $tf(t_i, d)$ is the term frequency of t_i term in the sentence, d and $tf(t_i, d)$ is number of times the t occurs in document d. The term frequency is calculated by using equation (3)

$$tf(t_i, d) = \log(f(t_i, d) \tag{3}$$

The inverse document frequency is utilized to find the infrequency of the term in the document. The inverse document frequency is calculated by using equation (4)

$$idf(t, D) = \log\left(\frac{N}{N_t} \in d\right) \tag{4}$$

Latent Dirichlet Allocation

After categorizing the words and indexing the terms in the sentence, the text summarization is carried out by using LDA for extracting the summary from classifiers such as Support Vector Machine(SVM), random forest, Gaussian naïve bayes. The text summarization is the method of extracting the useful words from the document of texts and it is done by using LDA. The LDA is an unsupervised probabilistic algorithm which isolate the high priority topics in a dataset as described by the keywords. The documents in the datasets are considered as random latent topics means which provides inferred than observed directly. The following steps are considered to model the document by LDA.

- The total number of words in the document are identified.
- Selected a particular topic for the document in a set of topics.
- A topic is chosen on the basis of document's multinomial distributions.
- Then a word is selected on the basis of topic's multinomial distributions.

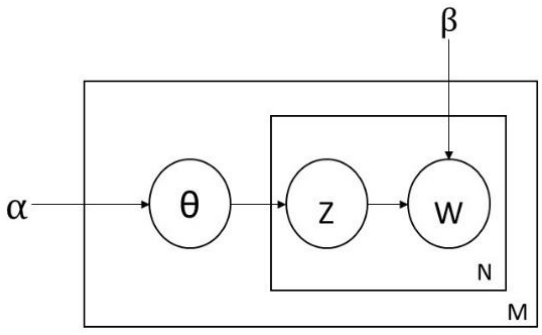

Figure 2. Graphical representation of LDA

The graphical representation of LDA model is shown in figure 2. There are N words in a document, V vocabulary words, K latent topics and M documents. Each word w of a documents d is associated with a hidden variables z which represent the latent topic. Variable z is sampled from a multinomial distribution with parameter T indicating the probability of latent topic. The density of T multinomial parameter is given by a dirichlet distribution with hyper parameter D. The $K * V$ parameter matrix $\beta = \{\beta_{kw}\}$ denotes the topics language model. The parameters of LDA$\{\alpha, \beta\}$ were estimated by increasing the marginal likelihood $p(w|\alpha, \beta)$ from a set of text documents $w = \{w_{dn}\}$.

$$\prod_{d=1}^{M} \int P(\theta_d|\alpha) \left[\prod_{n=1}^{N} \sum p(w_{dn}|z_{dn}, \beta) p(z_{dn}|\theta_d) \right] d\theta_d \quad (5)$$

where the marginalization is controlled by latent topic $z = \{z_{dn}\}$ and Dirichlet parameter θ_d. The variation inference is feasible to evaluate LDA parameters by a lower bound of (1) as a representative for optimization. By considering the variational inference factor where z and T are independent, a variational mode $q(\theta, z|\gamma, \phi)$ is obtained to optimize the true posterior probability $p(\theta, z|w, \alpha, \beta)$. To increase the lower bound the variation parameters are $\{\gamma, \phi\}$ and the LDA parameters $\{\alpha, \beta\}$ are calculated by equation (6)

$$\phi_{nk} \rightarrow \beta_{kw_n} \exp\{\Psi(\gamma_k) - \Psi(\sum_{j=1}^{k} \gamma j\} \quad (6)$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk} \quad (7)$$

$$\beta_{kw_n} \rightarrow \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dnk} w_{dn} \quad (8)$$

$$\alpha^{t+1} = \alpha^t - H_{LDA}(\alpha^t)^{-1} G_{LDA}(\alpha^t) \quad (9)$$

Where Ψ is the first derivative of log gamma function and t is the iteration index in decent algorithm. H_{LDA} and G_{LDA} is the hessian matrix and gradient vector of the lower bound by considering α respectively.

Classification Algorithms

The summary is extracted by using the classifiers such as SVM, Random Forest and Gaussian Naïve bayes which are explained in the following.

❖ SVM

SVM is supervised machine learning approach which is fast and dependable that utilizes the classification algorithms for 2 sets of problem in classification. After providing an SVM approach groups of labelled training data for all categories which are able to classify the new texts.

❖ Random Forest

Random forest is a supervised learning approach and it is utilized for classification as well as regression, it is most flexible and easy to use. Random forest makes decision trees on datasets that are collected randomly and provides prediction from every tree and gets the optimum solution.

❖ Gaussian naïve bayes

Gaussian naïve bayes is a variant of naïve bayes classifier which follows the gaussian normal distribution and supports the continuous data. Naïve bayes are set of supervised machine learning approach for classification on the basis of bayes theorem it is simple technique and includes high functionality.

IV. EXPERIMENTAL RESULT AND DISCUSSION

In this section, the experimental results of the proposed method are explained which effectively classifies the web pages based on visual and textual contents. The website is playing an important role by providing informations to users and is limited by fixed URL then displays the content of site as time-variant. The datasets such as URLs were randomly collected from the E-commerce webpage like Flipkart, Ebay and Myntra. The morphological analysis is carried out for the extracted information to identify the various form of words. The text summarization of the proposed method is carried out by using LDA. The system requirements of the proposed method were Random Access Memory(RAM) and intel. The performance metrics considered in the developed method were Accuracy, Precision, F-score, and Recall.

❖ Accuracy:

Accuracy is one of the important parameters for computing classification models. Generally, the accuracy is evaluated by the fraction of predictions. The definition of accuracy is given in Eq. (10),

$$Accuracy\ (\%) = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} \times 100 \quad (10)$$

❖ Precision:

To get the value of precision, the total number of correctly classified positive examples by the total number of predicted positive examples. The high precision indicates an example labelled as positive is indeed positive (Small number of False Positive (FP)). Precision is calculated by using Eq. (11).

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (11)$$

❖ Recall

The ratio of correctly predicted as fault-modules is defined as recall. The proportion of actual positives is correctly predicted by using recall, which is shown in Equation. (12)

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

❖ F1-Score

F1-Score measures accuracy of the model on a dataset, it combines precision and recall of the model. It is termed as the harmonic mean of the method's recall and precision. F1-Score is calculated by using equation (13).

$$F1 - Score = \frac{TP}{TP+1/2(FP+FN)} \quad (13)$$

Where,

TP =True Positive

TN =True Negative

FP = False Positive

FN = False Negative

Metrics	Weighted frequeny occurrence Without TFIDF			Weighted frequeny occurrence With TFIDF		
	Support Vector Machine	Random Forest Classifier	Gaussian Naive Bayes	Support Vector Machine	Random Forest Classifier	Gaussian Naive Bayes
Average Accuracy(%)	56	64	72	68	76	92
Average Precision(%)	31	64	73	46	76	92
Average Recall(%)	56	62	72	58	76	92
Avg.F1 Score(%)	40	60	69	45	76	92
MacroAverage Precision(%)	38	64	73	34	72	91
MacroAverage Recall(%)	50	60	64	50	72	91
MacroAvg.F1 Score(%)	36	60	64	40	72	91

Table 1. Performance comparison of weighted frequency occurrence without TFIDF and with TFIDF

Table 1 shows the comparison for the performance evaluation of weighted frequency occurrence without TFIDF and with TFIDF on the classifier algorithms such as Support Vector Machine, Random Forest and Gaussian Naive Bayes by considering the performance metrics like average accuracy, average precision, average recall, average F1-Score, macro average precision, macro average recall and macro average F1-Score. The figure 3-9 shows the comparison graph of weighted frequency occurrence without TFIDF and with TFIDF in terms of accuracy, precision, recall, F1-score, macro average precision, macro average recall, macro average F1-score.

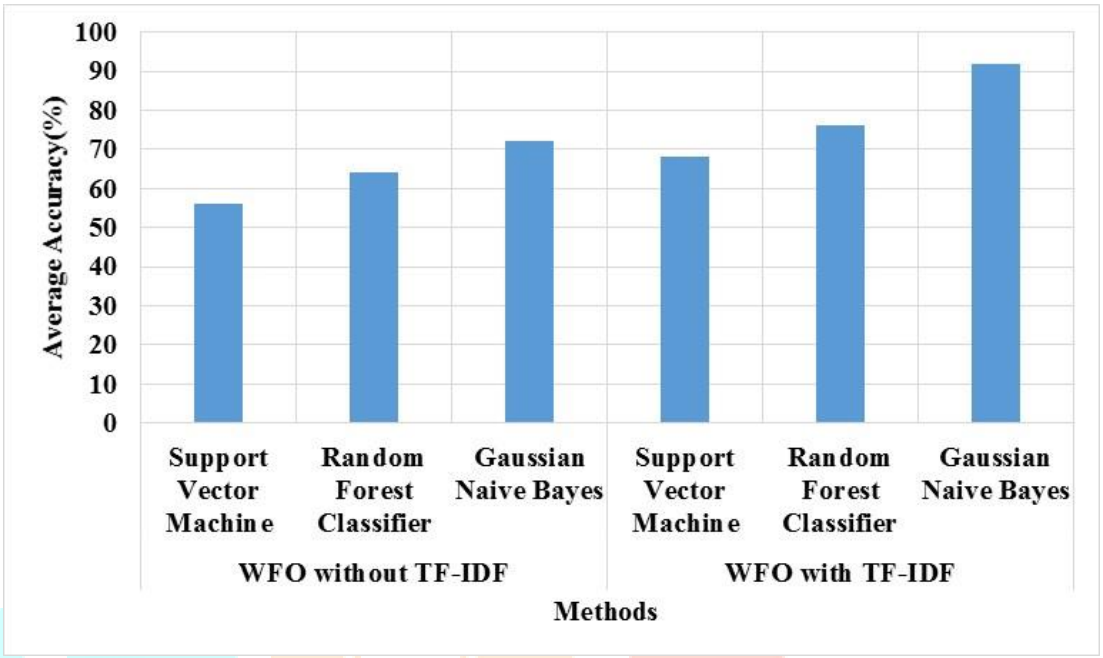

Figure 3.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Average Accuracy

Accuracy is the important parameter for computing the classification model and is explained in (1). The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of accuracy. The classification algorithm such as SVM, random forest and Gaussian naive bayes shows higher accuracy in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows accuracy of 92%, random forest of 76% and SVM of 68% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows accuracy of 72%, 64% and 56% in Gaussian naïve bayes, random forest and SVM which is shown in figure 3.

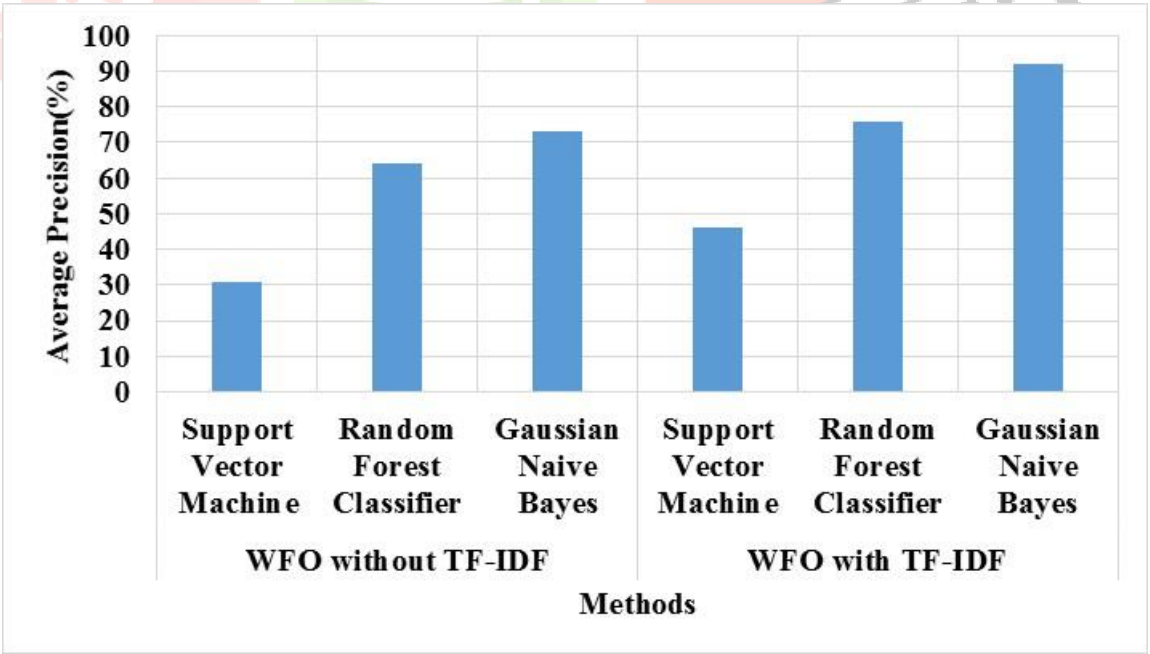

Figure 4. Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Average Precision

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of average precision rate. The classification algorithm such as SVM, random forest and Gaussian naive bayes shows higher average precision rate in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows average precision rate of 92%, random forest of 76% and SVM of 46% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows average precision rate of 73%, 64% and 31% in Gaussian naïve bayes, random forest and SVM which is shown in figure 4.

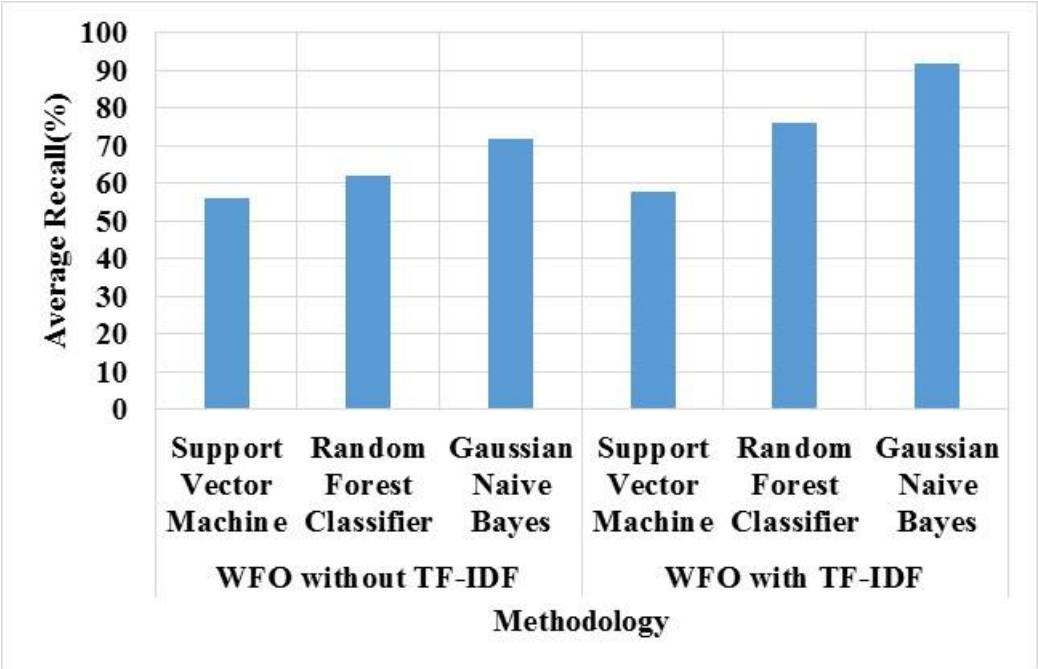

Figure 5.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Average Recall

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of average recall rate. The classification algorithm such as SVM, random forest and Gaussian naive bayes shows higher average recall rate in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows average recall rate of 92%, random forest of 76% and SVM of 58% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows average recall rate of 72%, 62% and 56% in Gaussian naïve bayes, random forest and SVM which is shown in figure 5.

Figure 6.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Average F1-Score

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of average F1-score. The classification algorithm such as SVM, random forest and Gaussian naive bayes shows higher average F1-score in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows average F1-score of 92%, random forest of 76% and SVM of 45% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows average F1-score of 69%, 60% and 40% in Gaussian naïve bayes, random forest and SVM which is shown in figure 6.

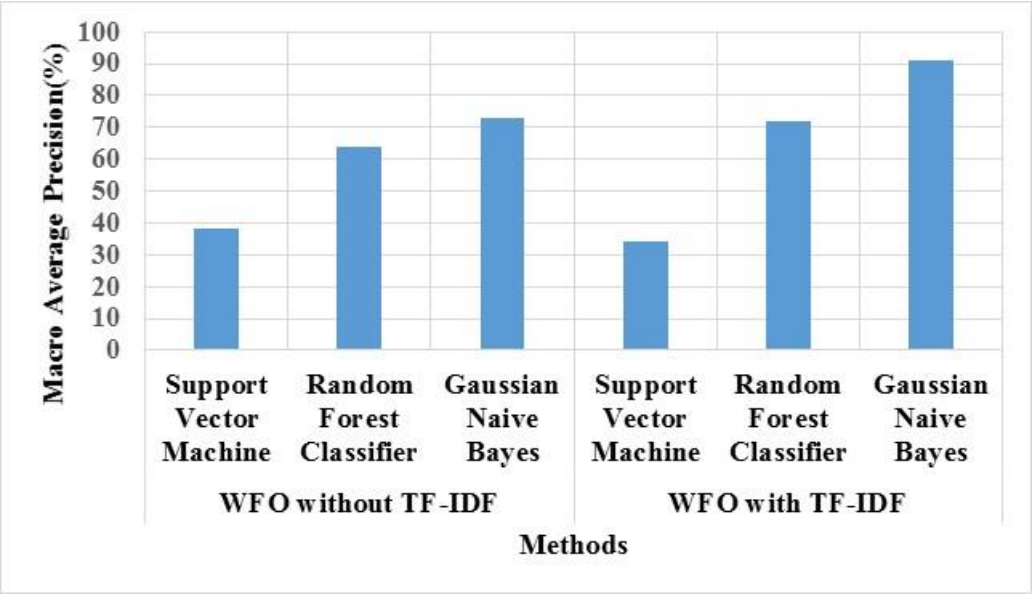

Figure 7.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Macro Average Precision

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of macro average precision. The classification algorithm such as random forest and Gaussian naive bayes shows higher macro average precision rate in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows macro average precision rate of 91%, random forest of 72% and SVM of 34% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows macro average precision rate of 73%, 64% and 38% in Gaussian naïve bayes, random forest and SVM which is shown in figure 7.

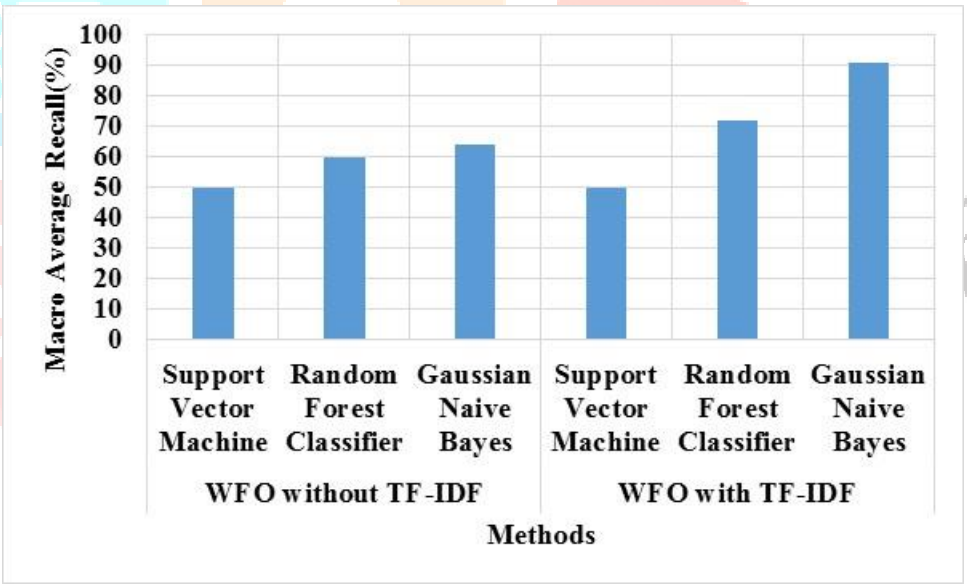

Figure 8. Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Macro Average Recall

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of macro average recall. The classification algorithm such as random forest and Gaussian naive bayes shows higher macro average precision rate in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows macro average recall rate of 91%, random forest of 72% and SVM of 50% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows macro average recall rate of 64%, 60% and 50% in Gaussian naïve bayes, random forest and SVM which is shown in figure 8.

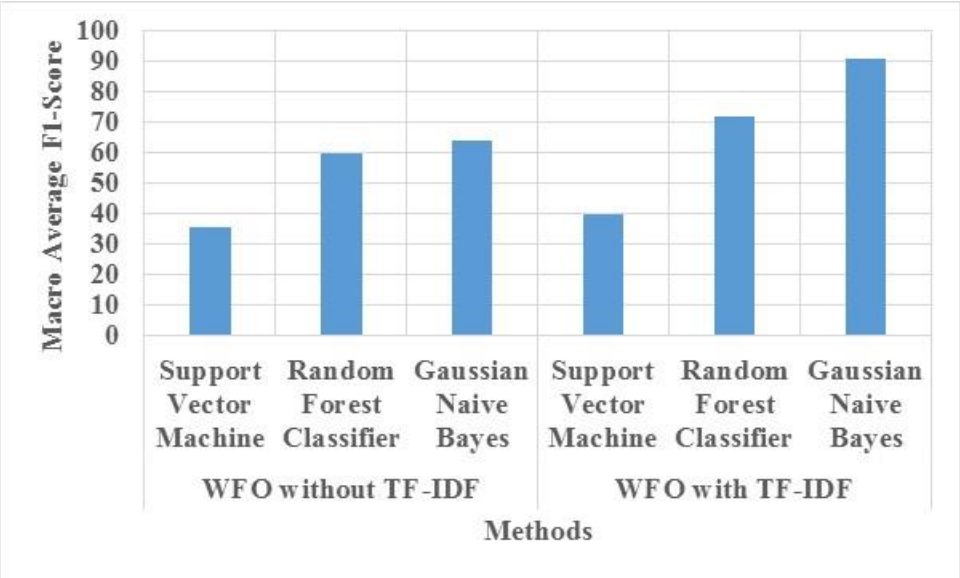

Figure 9.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of Macro Average F1-Score

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of macro average F1-score. The classification algorithm such as SVM, random forest and Gaussian naive bayes shows higher macro average F1-score in weighted frequency occurrence with TF-IDF. The Gaussian naive bayes shows macro average F1-score of 91%, random forest of 72% and SVM of 40% in weighted frequency occurrence with TF-IDF. Whereas, in weighted frequency occurrence without TF-IDF shows macro average F1-score of 64%, 60% and 36% in Gaussian naïve bayes, random forest and SVM which is shown in figure 9.

Rouge Metrics	Weighted frequeny occurrence Without TFIDF	Weighted frequeny occurrence With TFIDF
Precision(%)	92.00	97.00
Recall(%)	36.00	40.00
F1 Score(%)	28.00	34.00

Table 2. performance comparison of weighted frequency occurrence without TFIDF and with TFIDF in terms of rouge metrics.

Table 2 shows the comparison for the performance evaluation of weighted frequency occurrence without TFIDF and with TFIDF in terms of rouge metrics such as precision, recall and F1-score.

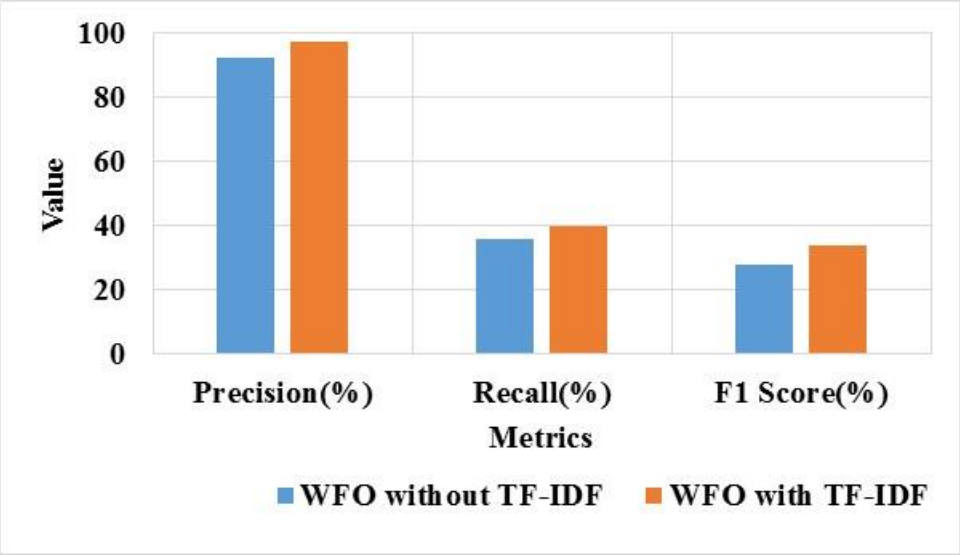

Figure 10.Comparison graph of weighted frequency occurrence without TF-IDF and without TF-IDF in terms of rouge metrics.

The comparison of weighted frequency occurrence without TF-IDF and with TF-IDF is evaluated in terms of rouge metrics such as precision, recall and F1-score. The WFO with TF-IDF shows the higher precision rate of 97%, recall rate of40% and F1-score of 34%. Whereas, the WFO without TF-IDF shows 92%, 36% and 28% in terms of precision, recall and F1-score.

V. CONCLUSION

In this paper, WFO method for image and content-based webpage information extraction has been proposed. The proposed system classifies the web pages based on visual contents and contexts information. The proposed methods are suitable for analysis of modern webpages where visual contents and textual contents have a dominant role. Initially, URLs were collected from the E-commerce webpage such as Flipkart, Ebay and Myntra. Then, the preprocessing process such as tokenization is carried out to breakdown the complex data. The WFO is calculated by using the frequency of words and maximum number of similar words present in the description. The calculated words are used in TF-IDF to categorize the words and to index the terms present in the sentence. Finally, text summarization of the proposed method is carried out by using LDA. The classifiers such as SVM, Random Forest and Gaussian Naïve bayes are used in the proposed method. The major contribution of the proposed method is the application of machine learning technique to the problem of visual content and textual content-based web page extraction. The experimental result shows that, Gaussian naïve bayes classifier shows higher performance of 92% in terms of accuracy, precision, recall and F1-Score.

REFERENCES

[1] Chen, M., 2019. Reducing Web Page Complexity to Facilitate Effective User Navigation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), pp.739-753.

[2] Jin, L., Feng, L., Liu, G. and Wang, C., 2017. Personal web revisitation by context and content keywords with relevance feedback. *IEEE Transactions on Knowledge and Data Engineering*, *29*(7), pp.1508-1521.

[3] Li, P., Li, T., Zhang, S., Li, Y., Tang, Y. and Jiang, Y., 2020. A semi-explicit short text retrieval method combining Wikipedia features. *Engineering Applications of Artificial Intelligence*, *94*, p.103809.

[4] Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, N. and Mahmood, W., 2019. The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. *IEEE Access*, *7*, pp.21662-21686.

[5] Khennak, I. and Drias, H., 2017. An accelerated PSO for query expansion in web information retrieval: application to medical dataset. *Applied Intelligence*, *47*(3), pp.793-808.

[6] Uzun, E., 2020. A Novel Web Scraping Approach Using the Additional Information Obtained from Web Pages. *IEEE Access*, *8*, pp.61726-61740.

[7] Gong, J., Zhang, H., Du, W., Li, H. and Wen, H., 2020. VB-PTC: Visual Block Multi-Record Text Extraction Based on Sensor Network Page Type Conversion. *IEEE Access*, *8*, pp.167900-167913.

[8] Vyas, K. and Frasincar, F., 2020. Determining the most representative image on a Web page. *Information Sciences*, *512*, pp.1234-1248.

[9] Xu, H., Huang, C. and Wang, D., 2019. Enhancing semantic image retrieval with limited labeled examples via deep learning. *Knowledge-Based Systems*, *163*, pp.252-266.

[10] Zhu, H., 2020. Massive-scale image retrieval based on deep visual feature representation. *Journal of Visual Communication and Image Representation*, *70*, p.102738.

[11] Bruni, R. and Bianchi, G., 2020. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications*, *142*, p.113001.

[12] Hu, M., Yang, Y., Shen, F., Zhang, L., Shen, H.T. and Li, X., 2017. Robust web image annotation via exploring multi-facet and structural knowledge. *IEEE Transactions on Image Processing*, *26*(10), pp.4871-4884.

[13] Lopez-Sanchez, D., Arrieta, A.G. and Corchado, J.M., 2019. Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing*, *338*, pp.418-431.

[14] Deepak, G. and Priyadarshini, J.S., 2018. Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Computers & Electrical Engineering*, *72*, pp.14-25.

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

[16] Huang, C., Xu, H., Xie, L., Zhu, J., Xu, C. and Tang, Y., 2018. Large-scale semantic web image retrieval using bimodal deep learning techniques. *Information Sciences*, *430*, pp.331-348.

[17] Bhopale, A.P. and Tiwari, A., 2020. Swarm Optimized Cluster Based Framework for Information Retrieval. *Expert Systems with Applications*, p.113441.