Covenant Journal of Informatics & Communication Technology. Vol. 6 No. 2, Dec., 2018

An Open Access Journal Available Online

A Review on Web Page Classification

Ayodeji Osanyin, Olufunke Oladipupo & Ibukun Afolabi

1Department of Computer and Information sciences, Covenant University, Ota, Nigeria osanyindeji@gmail.com, funke.oladipupo @covenantuniverity.edu.ng, ibukun.fatudimu@covenantuniversity.edu.ng

Abstract—With the increase in digital documents on the world wide web and an increase in the number of webpages and blogs which are common sources for providing users with news about current events, aggregating and categorizing information from these sources seems to be a daunting task as the volume of digital documents available online is growing exponentially. Although several benefits can accrue from the accurate classification of such documents into their respective categories such as providing tools that help people to find, filter and analyze digital information on the web amongst others. Accurate classification of these documents into their respective categories is dependent on the quality of training dataset which is dependent on the preprocessing techniques. Existing literature in this area of web page classification identified that better document representation techniques would reduce the training and testing time, improve the classification accuracy, precision and recall of classifier. In this paper, we give an overview of web page classification with an in-depth study of the web classification process, while at the same time making awareness of the need for an adequate document representation technique as this helps capture the semantics of document and-also contribute to reduce the problem of high dimensionality.

Keywords/Index Terms— Classification, Document representation, TF-IDF, Web Page classification, Word2Vec

1. Introduction

Aggregating and categorizing information from these sources seems to be a daunting task as the information on the World Wide Web is increasing every second at a very high rate due to the

influx of internet usage (Raj *et al.,* 2016). Automatic web classification / categorization is the main technology to achieve this.

Web page categorization which is also referred to Web Page Classification (WPC) is the process of assigning a web resource to one category or the other (Deri *et al.,* 2015). WPC problem can be sub-divided into two categories: the traditional manual method and the automated method of web page categorization. The traditional manual method is typically performed by experts who assigns web pages manually to the correct category, but this is impossible nowadays because of the influx of digital documents which will take a great deal of effort and time (Dey Sarkar *et al.,* 2014). While the former uses humans to achieve the categorization, the automatic method of WPC uses classification algorithms to determine the correct category in which the web pages belongs to automatically (Shibu *et al.,* 2010).

The former is tedious and time consuming, the latter reduces the large number of manpower, time needed for the classification, as well as resources (Dixit & Gupta, 2015).

Web classification is different from the standard text classification in some aspects: Traditional text classification is typically performed on structured documents which are stored in structured data stores such as relational databases and written with consistent styles which web collections do not possess (Qi & Davison, 2010; Abdelbadie *et al.,* 2013; AbdulHussien, 2017).

Web documents are semi-structured, formatted with the web markup language (HTML) which increases the rendering of the web pages to users. Also, web pages are linked together by hypertext within the same page or from one document to another (Qi & Davison, 2010). Several benefits can accrue from the accurate classification of documents into their respective categories such as providing resources that help the users to locate and retrieve the pertinent information amidst the vast resources on the web. Also news filtering, document routing and personalization of information on the web are additional advantages that can be harvested from web page classification.

According to (Mangai *et al.,* 2012), the commercial applications of web page categorization are as follows: most web directories owned by I.T giants such as Google, Yahoo and Microsoft bing are built, retained and extended by advanced WPC technologies (Huang *et al.,* 2004; Mangai *et al.,* 2012). Web page categorization are used to produce better search results from a search query. Searching for a particular resource proceeds by entering a keyword, and the search engine provides results related to the keyword, with WPC the search engine provides relevant and increased search results (Tsukada *et al.,* 2001). Advanced WPC techniques are used to improve the answers from a search result in a question and answering system (Cui *et al.,* 2004). Also, another very important application of WPC is web content filtering (Hammami *et al.,* 2003). Many WPC system have been proposed by several authors over the years, in which different approach have been formulated

to tackle the problem of the classifier performance (Kato & Goto, 2016). Amongst the notable machine learning algorithm which have been proposed by several authors in literature for web page classification include Naive Bayes, KNN, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision trees (DT) (Fatima & Srinivasu, 2017). The classification result of the web page classification system in achieving high result is dependent on making the pre-processed

document represent as much information as contained in the original document i.e. the pre-processing stage determines the quality of the results of the web classifier (Wang *et al.,* 2016). Also, the accuracy of most classification algorithms relies on the quality and size of training data which is dependent on the document representation technique (Dey Sarkar *et al.,* 2014).

This paper is a review paper which is intended in exploring the research question in the net section in a bid to achieve a systematic review of web page classification process, evaluate the document representation techniques and methodology used in web page classification.

2. Research Questions

The aim of this research is to answer the following research questions (RQs):

RQ1: What is the state of the art on WPC process?

The motivation for this question is to identify the current stages involved in the WPC process

RQ2: What are the Corpuses Used in web classification systems?

The purpose of this question is to discover the recent corpus or training data set used for web page classification

RQ3: What are current document representation techniques utilized in web classification systems?

The purpose of this question is to identify the gaps in the DR technique (semantic matching) utilized in web page classification

RQ4: What feature of the web page is used for web classification systems? The purpose of this question is to identify the main part of web page that is used for building the web page classification system

RQ5: What kind of methods are used for web page classification

The motivation for this question is to discover trends applied methodologies used in web page classification and thereby establish the state of the art methodologies

3. Methodology

The research questions are structured to express content of literature review particularly following the approach of (Webster & Watson, 2002) and of (Kitchenham, 2004). Scopus, IEEExplore, CiteSeerx, ACM library, Google Scholar were the main source for the publication due to their richness and relevance in content as regards to web page classification publications. The initial search keyword in Google scholar was ―web page classification‖ in the search bar, sorted by year and relevance. Then the search keyword was refined to consist of the following: ―web page categorization‖ ―feature selection techniques for web classification‖ ―document representation techniques for web page classification‖ ―web page classification process‖ ―semantic matching‖ ―Word2Vec for web page classification‖ ―automatic text categorization‖ ―topic models for web

page classification‖ ―comparison of document representation techniques‖ in the article title. 80% of the papers used were found in the Google Scholar database.

Inclusion criteria:

(1) Web page classification, Web page categorization, Document representation techniques, web page classification process, topic models for web page classification, LDA for web page classification are used to arrive at the search criteria and the major topics of the publications, (2) In a situation where several articles have reports that are similar, the latest publication is selected.

Exclusion criteria:

(1) Web classification using ontology based publications were excluded, because this research is focused on statistical techniques used in web page classification. (2) Publications that focused on general web classification using the multimedia content were excluded. (3) The contents of some online journal publications that could not be retrieved were also removed.

The necessary relevant criteria's highlighted above was the reference point for the articles and abstracts of the journal publications. In situations when details of the title and the abstract of the article don't match with the set of criteria, the whole content of the journal publication is examined, after which a decision for choice for either inclusion or exclusion is the made. The above highlighted procedure resulted in to 70 publications which was included in the next stages in the research process. These 70 publications were selected from a total of 85 which was retrieved before applying the inclusion criteria. The year of the publications selected ranged from 1999 to 2018.

3.1 RQ1: What is the state of the art on web page classification process?

This template was designed for two affiliations. According to (Fatima & Srinivasu, 2017), the web page classification system is divided in to several components as shown in Figure 1 below. The stages of the Web Page classification process includes: Creating a corpus of web pages, pre-processing / document representation, organization of the pre-processed pages, building the WPC model, obtaining a trained classifier, evaluating the classifier.

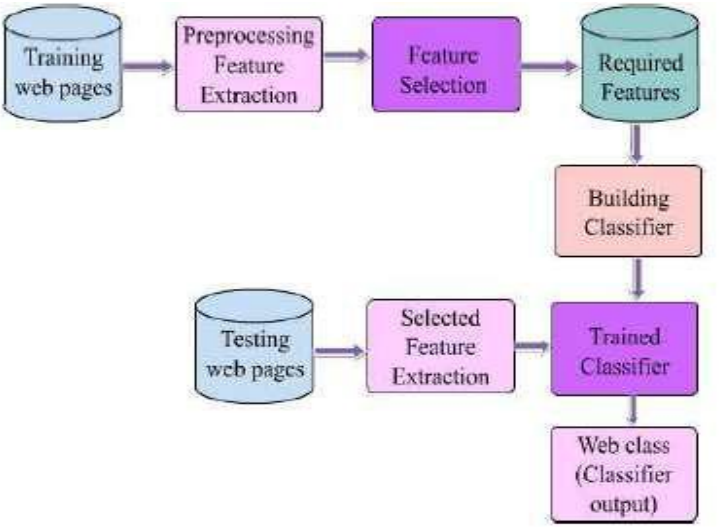

Figure 1. Web page classification process

A. Corpus or Web Pages Training Dataset

The first stage in the web classification process proceeds with extracting the main contents of the webpage along with other web page elements such as Internal and external hyperlinks, Metadata, Flash animation, Java script, Video Clips, Embedded objects, advertisement, Google ad-sense (Deri, Martinelli, Sartiano & Sideri, 2015). The extracted web contents are used in creating a corpus of labeled web pages i.e. training web pages which would be utilized by the classifier to building the learning system (Qi & Davison, 2009).

B. Pre-processing/Document Representation

The next stage in the web page classification process is the pre-processing stage also known as Document Representation (DR) or dimensionality reduction in this context (Mangai, Kothari & Kumar, 2012). This stage can also be further sub-divided into Feature Extraction (FE) and Feature Selection (FS) (Baharudin, Lee & Khan, 2010). FE process begins by extracting the raw content of the pages and discard HTML tags and other WWW contents. Web page document are characterized by high dimensionality, the first technique to reduce this high dimensionality is FE (Shibu, Vishwakarma & Bhargava, 2010; Raj, Francis & Benadit, 2016).

Then FE process continues by breaking down text into small chunks known as token which can either be a phrase, word or symbols in a process known as tokenization. After tokenization, then the tokens are reduced to their root or inflectional words know as stemming or lemmatization. Then lowercase conversion and filtering out of stop words (They are generally regarded as 'functional words' which do not carry meaning such as ―the‖, ―a‖, ―and'') (Fatima & Srinivasu, 2017).

The feature selection stage precedes after feature extraction. This stage involves constructing a vector matrix of the web document which is aimed at improving the accuracy of the web

classifier. The basic aim of feature selection is to select the most important features that would represent the whole document (Alamelu Mangai *et al.,* 2010). Also with the inherent characteristics of web document which is high dimensional datasets, FS is used to reduce the space the original high dimensional space to a lower dimensional space which helps to increase the overall accuracy of the classifier and efficiency. Feature selection approaches can be broadly classified as filter, wrapper, and embedded.

The most generic of all the approaches is the filter approach which is the independent of the classification being utilized (Dey Sarkar, Goswami, Agarwal & Aktar, 2014). The filter approach uses metrics such as mutual information, correlation, entropy and so on, which analyzes general the general structure of the dataset and selects the optimal feature set (Talavera, 2005). The filter approach is a straight forward method and easier to work out than the other (embedded and wrapper) approaches (Kojadinovic & Wottka, 2000).

However, it is to be noted that wrapper and embedded methods often outperform filter in real data scenarios (Alamelu Mangai *et al.,* 2010). But in former (embedded approach), the algorithm is designed to embed the FS together with the objective function. Examples of embedded approach are DT, LASSO, 1-N SVM and so on. While in the later (wrapper approach), works on the basis of several combinations of the whole data for training and testing, which is usually an exhaustive search for the target function that learns the best feature set for the dataset. Metrics such as classification accuracy are used for selecting the optimal feature set. A major drawback of this approach is that, it is computationally expensive because of the brute force approach (Dey Sarkar *et al.,* 2014).

In contrast to the approaches discussed earlier, that selects the optimal feature set from the set of features, other techniques try to transform the original high dimensional feature matrix in to a lower dimensional matrix. This effectively helps to determine the semantics of a document and also, the main concepts in the document (Said, 2007; Qi & Davison, 2009; Li, Xia, Zong & Huang, 2009). Also, the above approaches cannot infer the inter or intra document statistical structure of the corpus (Biro, Benczur, Szabo, Maguitman, 2008). Such methods include: bag of words model TF-IDF (Ayyasamy *et al.,* 2010; Wang *et al.,* 2013; Sartiano & Sideri, 2015; Weiping & Chunxia, 2015; Moiseev, 2016; Raj *et al.,* 2016; Deri *et al.,* 2015; Fatima and Srinivasan, 2017), Latent Semantic Indexing (LSI) (Deerwester *et al.,* 1990; Chen & Hsieh, 2006; Biro *et al.,* 2008), Probabilistic Latent Semantic Indexing (PLSI) [34], Word2Vec (Lilleberg *et al.,* 2015; Wang *et al.,* 2016), Latent Dirichlet Allocation (LDA) (Sriurai *et al.,* 2010; Špeh *et al.,* 2013; Wang *et al.,* 2016).

Each technique has its own pros and cons. Lots of discussions are ongoing in the pre-processing and document representation stage of the WPC system. Document representation is very crucial stage in the web page classification process as irrelevant and noisy features

in the data set will impact badly on the performance of the classifier in terms of its accuracy, speed and reducing overfitting issues (Alamelu Mangai *et al.,* 2010).

Also this stage has gain more attention recently than any other component of the WPC as good dimensionality reduction will improve the learning capabilities of the classifier and good storage capabilities (Ayyasamy *et al.,* 2010; Azam & Yao, 2012; Dey Sarkar *et al.,* 2014; Lilleberg *et al.,* 2015).

C. Obtaining the Required Features

The next stage after pre-processing stage is to gather the required feature set for classification which is usually achieved by creating matrix representation of the document vectors which would be fed to the classifier (Alamelu Mangai *et al.,* 2010).

D. Building the WPC Model

After gathering the required features, the next stage is to build the WPC model using a classification algorithm with the selected features as the input data set. Several machine learning algorithm have been used for the building the model of the WPC system systems such as KNN (Miao *et al.,* 2009; Bang *et al.,* 2010; Karima *et al.,* 2012), Support Vector Machine (SVM) (Chen & Hsieh, 2006; Sriurai, Meesad & Haruechaiyasak, 2010; Patil & Pawar, 2012; Wang, Chen, Jia & Zhou, 2013; Lilleberg, Zhu & Zhang, 2015; Fatima & Srinivasu, 2017), Naïve Bayes (Dey Sarkar et, al., 2014; Raj, Francis & Benadit, 2016), Decision trees (DT) (Kim *et al.,* 2001), Deep Learning (Kato & Goto, 2016), Weighted Voting of Feature Intervals known (WVFI) (Mangai *et al.,* 2012a; Mangai *et al.,* 2012b), Artificial Neural Network (ANN) (Ruiz & Srinivasan, 1998; Yu *et al.,* 2008) and so on. After training the classifier, the model obtained is thereafter utilized to automatically categorize new web resources to the appropriate category.

Several authors have argued about the best ML technique for web page classification but literature has shown that the accuracy, generalization capabilities of any ML technique depends on the training data set i.e. choice of the techniques used in the preprocessing stage have an overall effect on quality of the classifier (Biro *et al.,* 2008; Karima *et al.,*; 2012; Dey Sarkar et, al., 2014; Wang, Ma & Zhang, 2016; Chao & Sirmorya, 2016; Singh *et al.,* 2017)

E. Evaluating the Classifier

To test the performance of the web page classifier, some evaluation metrics are utilized to do this. A confusion matrix is one of the widely used metric for evaluating the performance, which is shown in table 1 below. A confusion matrix is a table that showcases the correct label of a category again the predicted label of a category. In the table, the total number of true positive classification is represented by ―i‖, while that of false positive classification is denoted by ―j‖. Also, the number of false negative classifications are denoted by ―k‖, while that of true negative classification are denoted by ―l‖. For a classifier to be of optimal performance, both j and k must be zero (Jindal *et al.,* 2015).

Table 1: Confusion Matrix

		Predicted Class	
		T_n	Not T_n
Actual Class	T_n	I	j
	Not T_n	K	L

The accuracy of the classifier can be calculate from the confusion matrix using the value obtained from this calculation: (i+j) / (i+j+k+l). Other metrics used for evaluating the performance of web page classification problems are known as Precision (Pr) and Recall (Re). The value of recall is calculated as i/(i + k) which is the total proportion of dataset in category Tn that are correctly predicted has been in that class. While the whole ratio of dataset which are correctly predicted to belong to that category Tn which belongs to that category. At every point of recall, a precision is associated with it. Most times, it is standard practice to combine both Pr and Re as a standalone metric called F1 score which is calculated from the computation of (Jindal *et al.*, 2015).

From the review above, it is has been highlighted that the web page classification process proceeds with the creation of corpus, pre-processing/document representation, obtaining the required features, building the WPC model and finally evaluating the classifier

3.2 RQ2: What are the corpuses used in web page classification?

To delve in to this question, we look in to reviewing the datasets utilized by several authors in building web page classification systems. Dataset utilized for web page classification include: Reuters datasets, which is a dataset created from Reuter's newswire and it contains 118 different category of news. Web Kb is another well-known dataset which has been utilized in several text classification problems. It is freely distributed online and it contains about 1065 web pages which is categorized in to two categories. Also the 20Newsgroups dataset is another popular dataset that is utilized for text categorization. It contains 20 categories of news items which is made up of 18846 different news in different categories. Another popular dataset is the yahoo news dataset which contains user's activities of yahoo websites and applications such as sports, finance and real estate (Wang *et al.*, 2016). Another corpus that is being utilized as training data set is the Imdb dataset which contains 1094 movie scripts downloaded from the Internet Movie Script Database (IMSDB) in HTML format. The movie scripts in this dataset are American Hollywood movies released from 1935 to 2015. The distribution of the genres of the movie in the corpus are drama, thriller, comedy, action, crime, romance, adventure, sci-fi, horror. Also, SOS dataset are frequently utilized which contains several categories of articles such as history, language studies, music, religion and so on, summing up to 4,625.

Figure 2 reveals the most important dataset utilized by several authors in web page classification systems. From the chart it is shown that Reuter's

dataset is still the most utilized dataset by several authors.

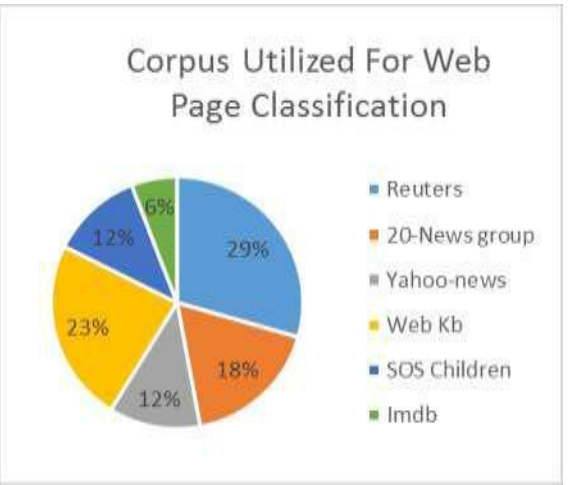

Figure 2. Corpuses utlilzed for Web Page Classification

RQ3: What are current document representation techniques utilized in web page classification?

According to Google index the amount of web resources available online is over 130 trillion pages and its growing at a rapid rate as due to fact that new users are added to the existing users every day. Retrieving information as soon as possible from this web documents is becoming necessary for many real life application (Azam & Yao, 2012). The accuracy and generalization capabilities of the classifier in assigning a web page to its correct category is heavily dependent on the document representation (Wang *et al.,* 2016). Several authors have applied various DR techniques to improve the quality of the input dataset which inherently will increase the general performance of the WPC system. Each technique is fraught by one challenge or the other. Some of them are highlighted below:

Term Frequency-Inverse Document Frequency (TFIDF)

This is the traditional and most popular method for document representation that is often used in information retrieval. It is a model that measures the importance of a word across a document. It weighs the important words increasingly based on how frequently they appear in the document but decreases the weight proportionally as it occurs in other documents. The TF-IDF weighting function is shown below:

$$W_{i,j} = tf_{i,j} * idf_{j} = = tf_{i,j} * \log_2\left[\frac{N}{df}\right]$$

The Term Frequency, $_{i, j}$ measures the no of occurrences of a word in a document:

$$tf_{i,j} = \frac{Number\ of\ times\ word\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

The Inverse Document Frequency, $_{j}$, measures the importance of a word by reducing the word's weighting score if it frequently occurs in other documents

$$idf_{i} = \log_2\left(\frac{Total\ number\ of\ document}{Number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

TF-IDF can represent a document well by removing stop words from the documents. Some of the drawbacks of tf-idf are that it does not capture semantic similarity, does not respect word order and it is an unordered collection of words.

Latent Semantic Indexing (LSI)

Another popular method used in information retrieval method which utilizes linear algebra index technique to tackle the sparse matrix produced by TF-IDF methods is referred to as Latent Semantic Indexing (LSI) (Deerwester, *et al.,* 1990). LSI uses a vector model to build a matrix of word co-occurrences. It identifies the position on a vector space where each term and a document in a collection are. It works on the assumption that groups of words are semantically related will cluster together (Landauer *et al.,* 1997). To create a low dimensional representation of the document, it utilizes SVD algorithm on the sparse bag of words matrix, to create a denser matrix that approximately models the original document. It composes frequencies of terms as a term-document matrix. LSI was used to solve the synonym and polysemy problem of TF-IDF.

However, a major drawback of LSI are that, it does not capture multiple meanings of a word and it does not respect word order (Zhang *et al.,* 2011). Also, LSA models a document as a Gaussian distribution while in most situation a Poisson distribution is observed and the resulting dimensions might be difficult to interpret (Biro *et al.,* 2008).

Probabilistic Latent Semantic Indexing (PLSI)

To overcome some of the afore-mentioned problems with LSI, (Hofmann, 1999) proposed a more sound approach referred to as Probabilistic Latent Semantic Indexing (PLSA), which uses a generative approach for enhancing the capabilities of latent semantic indexing (LSI). The model obtained by PLSI is usually a probabilistic co-occurrence of words as a mixture of generative words. It uses EM Algorithm for its learning (Oneata, 1999). PLSI is usually viewed as a more sound method as it provides a probabilistic interpretation, whereas LSI achieves the factorization by using only mathematical foundations (more precisely, LSI uses the singular value decomposition method) (Batra & Bawa, 2010). Also, PLSA deals with synonyms and polysemy words by taking a deeper look at different forms of words and meanings. The core foundation of probabilistic latent semantic analysis are statistical models.

The introduction of PLSA shows promising results but it has two major drawbacks which are: the hyper-parameters are linear in nature while real life web documents are not, which impacts on predicting of new documents (Biro *et al.,* 2008).

N-Gram Model

Another popular method for document representation is the N-gram model. It is based on the assumption that any given word can be predicted based on the probability of its proceeding n-1 word, where n = 1, 2, 3.....x, x is a whole number. If n = 1, it is referred to as a unigram model, when n = 2, it is a bigram, 3 is a trigram. N-gram approach to feature representation converts a corpus of text in to the corresponding feature vector by taking

record of the n-gram frequency counts which will serve as input vector to the classifier (Cianflone & Kosseim, 2017). The N-gram model can be of two forms. The first is referred to as character n-grams model, it rest on the assumption that sequence of unique occurring letters in a corpus while the second refers to as word n-gram model which relies on sequence of unique and occurring words in a corpus. Word N-gram model outperforms character n-gram model in many real word applications (Giannakopoulos *et al.,* 2012). According to (Elberrichi & Aljohar, 2007), some of the major strengths of N-GRAM are: No need to performing word segmentation. Capturing of root words automatically by the model. All languages are independent of each other. It has a low tolerance with distortion of words and mistakes usually made with spellings. In addition, no dictionary or language specific techniques are needed (Wei *et al.,* 2009).

N-GRAM suffers from data sparcity and high dimensionality (Mikolov *et al.,* 2013).

Latent Dirichlet Analysis (LDA)

Latent dirichlet analysis also known as LDA is a probabilistic topic model that generates what is referred to as latent topics based on the occurrence of a word in a text corpus or documents (Blei *et al.,* 2003). It assumes documents are a blend of several topics and that each word in the document can be grouped under the document's topics. LDA is typically handy is situations where there is need to find accurate mixture of topics within a given document. LDA is an unsupervised language model that transforms words from bag of words counts into continuous representative matrix. According to (Blei *et al.,* 2003), LDA works with the assumption that the generative process for each document in a collection of documents D is as follows:

1. $Choose\ N \sim Poisson\ (\varepsilon)$
2. $Choose\ \emptyset \sim Dir\ (\alpha)$
3. $For\ each\ of\ the\ N\ words\ W_n$
4. $(a) Choose\ Z_n\ Multinomial\ (\emptyset)$
5. $(b) Choose\ a\ word\ W_n\ from\ p(W_n|Z_n,\beta) a\ multinomial\ probability$
6. $Condition\ on\ the\ topic\ Z_n$

A major drawback of LDA is that improper calibration of these parameters could lead to sub-optimal results. Also, LDA uses an unsupervised learning function which depends on words in the corpus which will determine the matching degree and thus will suffer from vocabulary mismatching problem (Dit *et al.,* 2013).

Word2vec

This is a neural network language model that can learn word embedding's. The 2 main architectures are CBOW (Continuous-Bag-of-word) and Skip-gram (Continuous-Skipgram Model). The first architecture tries to predict words from the context of words while skip-gram tries to predict the context from the words. In the CBOW model each input vector u (i) is a column in the Matrix U. The CBOW model predicts a word u (i) utilizing the context u (i − n)... u (i − 1), u (i + 1)..., u (i + n), while the Skip-gram model predicts each word in the context utilizing the word u (i). The Word2Vec framework aims at predicting the context of word or word based on their context. The word embedding's are learned through maximizing the objective function. With these word embedding's it can capture

distributed representations of text to capture similarities among concepts (Mikolov *et al.,* 2013) which is one of the major advantages of Word2Vec. However, a major drawback of word2Vec is that it does not model the global relationship between documents to topics (Wang *et al.,* 2016).

According to (Singh *et al.,* 2017), many new hybrid techniques have been formulated by several authors to harness the strength of each of the technique highlighted above for adequate preprocessing of the input data: LSI and TF-IDF (Chen & Hsieh, 2006; Zhang *et al.,* 2011), N-Gram and TF-IDF (Karima *et al.,* 2012), Word2Vec and TF-IDF (Lilleberg *et al.,* 2015), Word2Vec and LDA (Wang *et al.,* 2016), TF-IDF and firefly Algorithm (Raj *et al.,* 2016; Ma *et al.,* 2016), TF-IDF and K-means clustering (Milios *et al.,* 2006; Dey Sarkar *et al.,* 2014), LDA and TF-IDF (Sriurai *et al.,* 2010), Doc2Vec and Affinity propagation (Ma *et al.,* 2016).

Figure 3 below shows that the dominant document representation technique utilized by most researchers is the bags of words model TF-IDF followed by LDA then Word2Vec. The chart also shows that the hybrid techniques are gradually becoming popular.

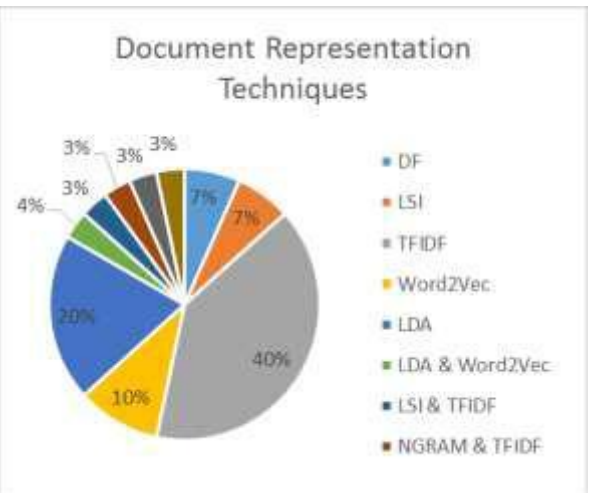

Figure 3: Document Representation Techniques

According to literature, there are many document representation techniques used for the preprocessing stage of the WPC and bag of words model (Bow) TF-IDF is still the most used DR technique

3.4 RQ4: What feature of the web page is used for web classification? To answer this question, we focus on reviewing the feature utilized in creating the training dataset for the WPC system. This is because web pages are semi-structured with HTML tags which is made of several parts such as the page content, Meta tags, links & URL and HTML structure. According to (Shibu *et al.,* 2010), building WPC system using the page content involves utilizing the content of the web page to determine the category in which the web page belongs

to. For Meta tags, the WPC system rely solely on attributes of Meta tags i.e. (<META name=‖Keywords‖> and <META name=‖description‖). For links and URL, this method is dependent on exploiting the contexts surrounding a link in an HTML document to extract useful information. HTML structures approach exploits both the content, html structure, images, placement of links contained in the page for building the WPC system. Figure 4 below shows that page content is the most used feature utilized for building the WPC system.

Figure 4: Web Page Features Utilized for WPC System

RQ5: What kind of methods are used for web page classification

According to (Xu *et al.*, 2011), a URL-based web page categorization using n-grams as the DR technique was proposed. Most updates users sends on social media like twitter, Facebook and so on contain links to email and webpages, there URL can be a means of categorization this information. Recently the influx of multimedia content on the web such as videos and images makes categorization by web pages a cumbersome process. In their work, they use n-Gram Language Model (LM) to classify textual data using the URL links of the web pages. The proposed method was applied to three datasets (webKb, DMOZ and GVO datasets). Results obtained showed an increase in the F1 measure for their method when compared with earlier methods

In the work of (Dey Sarkar *et al.*, 2014), they tried to solve the document representation and feature selection problem in web page classification. The methodology employed involved using chi-square metric to select the important words. The selected words are represented by their occurrence in various documents by simply taking the transpose of the term document matrix. K-means clustering is used to prune the

feature space further to reduce the dimensionality of the term document matrix. Naive Bayes classifier is then fitted against the document to classify the document in to its appropriate class. Their proposed methods was applied to thirteen datasets and experimental results shows that their method outperforms other earlier feature selection techniques. A major gap identified in their work is that document representation technique utilized was Document Frequency (DF) which is bag of words model and it is an inherent problem of not capturing the semantic similarity and word order of the document (Singh *et al.,* 2017).

Deri *et al.,* (2015), formulated a classification tool for TLD (top level domain). The methodology preceded by creating a custom made crawler that is able to download web pages starting from the index page, removing non HTML web pages and automatically discarding irrelevant pages such as about us page and so on. Then using python NLTK processor to perform traditional text processing such as lemmatization, stop words removal and so on. They used a bag of words model (tf-idf) to construct the term document matrix. Then the terms were then trained using classification algorithm (Naive Bayes and SVM). Results obtained shows that Naive Bayes classifier performs better than the SVM algorithm using Precision, recall and F1 score. A major gap identified in their work is that the document representation technique used was TF-IDF which is a bag of words model which does not capture semantic similarity and word order of the document being transformed (Wang *et al.,* 2017).

According to (Lilleberg *et al.,* 2015), they applied neural network model (Word2Vec) with bags of words model (tf-idf) to solve the document representation problem of web classification. Accurate Representation of documents affect the correct classification or categorization of new documents. To solve the document representation problem, they created a hybrid of Word2Vec weighted with tf-idf with stop words to correctly represent the feature vectors of a document. The proposed method was applied to 20 newsgroup text dataset. Results obtained shows that there proposed method outperforms tf-idf with/without stops words and word2vec with/without stop words. A major drawback with their work is that, stops words increase the dimensionality of the feature vectors which impacts badly on the classification accuracy and computational burden (Wang *et al.,* 2016). Also the classification algorithm used was a linear SVM, other kernels such as string and RBF kernels could produce better results (Nayak *et al.,* 2015).

In the works of (Raj *et al.,* 2016), they proposed a method to automatically classify web pages into different categories via three stages, which are: FE, information learning and classification. In the methodology adopted, term document matrix is created using tf-idf, then the terms are used to extract object based features. Decision tree algorithm is then used to generate rules from the features set. The rules extracted are then used as input in to the hybrid of optimal firefly algorithm based Naive Bayes Classifier (FA-NBC). The proposed method was

applied to WebKB datasets. Experimental results shows that their proposed method outperforms earlier methods such as KNN. Drawbacks identified in their work include: using tf-idf to construct the term document matrix does not capture any semantic similarity or form of grammatical analysis (Wang *et al.,* 2016).

Moiseev (2016), proposed a method to analyze and categorize e-commerce websites automatically. In their methodology, e-commerce website were crawled, text preprocessing and the terms of the document were derived using tf-idf. The proposed method was applied 1312 e-commerce and 1077 non e-commerce web site, preprocessing of the webpages, term weighted with tf-idf and classified using SVM. Experimental results shows that the produced method outperforms pure TF-IDF. Also the results shows a substantial increase in the accuracy of the classifier. A major gap identified in their word is that bag of words model like TF-IDF does not capture semantic similarity and respect word order of the document being represented (Singh, Devi & Mahanta, 2017).

In the works of (Wang *et al.*, 2016), they proposed the use of a hybrid strategy that consist of Latent Dirichlet Allocation (LDA) and Word2Vec for document representation. Word2Vec create a vector representation of the document which shows the semantic relationship between the words of the document. Euclidean distance was used to measure and interpret similarity between document and topic in sparse space. Their methods was applied to 20 News group data using SVM classifier. Results obtained shows that their proposed methods outperforms earlier methods such as TF-IDF+ SVM, Word2Vec + SVM, LDA + SVM. One of the major drawback of their method is that hyper-parameter tuning of LDA parameters i.e. # of topics, could produce unsatisfactory results as most of the parameters for the LDA are imported from natural language community (Dit *et al.,* 2013).

Based on the review conducted, several authors have proposed myriads of methods of improving the accuracy of the WPC system at the pre-processing or document representation stage, therefore showing this stage is still open to more research.

Observations

Representation of the input data (DR) is a crucial issue in web page classification and text classification systems at large. The performance of an algorithm is determined by the function of the input data available (Oyelade et al., 2010). Several feature selection techniques have been proposed to solve the issue of semantic matching of unstructured data, but are marred with one issue or the other. Recently, there has been an increase in the use of SVM and KNN for text classification (Khan et al., 2010; Jindal et. al, 2015). Also from extant literature, SVM, KNN and Naïve bayes are one of the most widely used ML algorithm for text classification (Joachims, 1998; Kwon & Lee, 2000; Asirvatham & Ravi, 2001; Sun et al., 2002; Khan et al., 2010; Sriurai et al., 2010; Krestel, 2012; Mangai et al., 2013; Lin & Wang, 2014; Lilleberg et al., 2015; Dixit & Gupta, 2015; Raj et al., 2016).

In the work of (Dey Sarkar et al., 2014), they decided to investigate this issue and

compared SVM, KNN and Naïve Bayes on text classification tasks. Results obtained shows that SVM was not a clear winner, despite quite good overall performance. If a suitable pre-processing is applied to KNN and Naïve Bayes theory, these algorithms will achieve very good results and scales up to the performance of SVM. In light of this, there is need for an adequate document representation technique to retrieve the semantics of a web document. Optimized document representation techniques such as hybridizing neural network language models (Word2Vec) and topic model (LDA) or Word2Vec and TF-Idf with optimizing the parameters of LDA with search algorithms (such as GA) will provide better semantics of the document in WPC. This hybrid approaches has shown to perform better (obtain the semantic features) by harnessing the strength of the individual technique in the arrangement Word2Vec and LDA (Wang et al., 2016) or Word2Vec and TF-Idf (Lilleberg et al., 2015]. Also, proper calibration of the parameters of LDA with a search algorithm would produce better latent topics across words in a document (Dit et al., 2013).

Conclusion

In this paper, we gave an overview of web page classification system. Different application areas and an in-depth analysis of the web page classification process were looked into. Analysis of state of art techniques for feature selection techniques used in WPC was looked in to with a view to identify challenges fraught by each one. Also related works in the areas of WPC was reviewed to identify the latest works in this domain. It clearly shows that document representation phase is one of the areas that are receiving interest by researchers. Most currently used methods of document representation are Vector Space Model (VSM), Probabilistic Topic Model and Statistical Language Models and Neural network language models [47]. The chosen document representation technique have a direct impact on the classification results as it captures the semantics of document and also contribute to reduce the problem of high dimensionality. Combining different DR technique are new areas of research because each technique perform differently depending on the dataset. Future work in WPC should focus on improving the semantic relationship of web document by hybridizing difference DR technique which will inherently improve the classification result. Also, ontology based techniques can be used to capture the real semantics of unstructured text (Daramola et al., 2013).

References

Abdelbadie B., Abdellah I., & Mohammed B. (2013). Web Classification Approach Using Reduced Vector Representation Model Based On Html Tags. *Journal of Theoretical and Applied Information Technology*, 55(1).

AbdulHussien, A. A. (2017). Comparison of Machine Learning Algorithms to Classify

Web Pages. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(11), 205-209.

Alamelu Mangai, J., Santhosh Kumar, V., & Sugumaran, V. (2010). Recent Research in Web Page Classification–A Review. *International Journal of Computer Engineering & Technology (IJCET)*, 1(1), 112-122.

Asirvatham, A. P., & Ravi, K. K. (2001). Web page categorization based on document structure. *Centre for Visual Information Technology*.

Ayyasamy, R. K., Tahayna, B., Alhashmi, S., Eu-gene, S., & Egerton, S. (2010, May). Mining Wikipedia knowledge to improve document indexing and classification. In Information *Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on IEEE*, (pp. 806-809).

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39(5), 4760-4768.

Bang, S. L., Yang, J. D., & Yang, H. J. (2006). Hierarchical document categorization with k-NN and concept-based thesauri. Information processing & management, 42(2), 387-406.

Batra, S., & Bawa, S. (2010). Using lsi and its variants in text classification. Advanced Techniques in Computing Sciences and Software Engineering, 313-316.

Biro, I., Benczur, A., Szabo, J., & Maguitman, A. (2008, October). A comparative analysis of latent variable models for web page classification. In Latin American Web Conference, 2008. LA-WEB'08. (pp. 23-28). IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Chao, B., & Sirmorya, A. (2016). Automated Movie Genre Classification with LDA-based Topic Modeling. *International Journal of Computer Applications*, 145(13).

Chen, R. C., & Hsieh, C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications*, 31(2), 427-435.

Cianflone, A., & Kosseim, L. (2017). N-gram and Neural Language Models for Discriminating Similar Languages. arXiv preprint arXiv:1708.03421.

Cui, H., Kan, M. Y., Chua, T. S., & Xiao, J. (2004, July). A comparative study on sentence retrieval for definitional question answering. *In SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)* (pp. 383-390).

Daramola, O., Afolabi, I. T., Akinyemi, I., & Oladipupo, O. O. (2013). Using ontology-based information extraction for subject-based auto-grading.

ResearchGate, 373-378.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391.

Deri, L., Martinelli, M., Sartiano, D., & Sideri, L. (2015, November). Large scale web-content classification. *In Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference* on (Vol. 1, pp. 545-554). IEEE.

Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014). A Novel Feature Selection Technique for Text Classification Using Naïve Bayes. *International Scholarly Research Notices*, 2014.

Dit, B., Panichella, A., Moritz, E., Oliveto, R., Di Penta, M., Poshyvanyk, D., & De Lucia, A. (2013, May). Configuring topic models for software engineering tasks in tracelab. *In Traceability in Emerging Forms of Software Engineering (TEFSE), 2013 International Workshop on* (pp. 105-109). IEEE.

Dixit, S., & Gupta, R. K. (2015). Layered Approach to Classify Web Pages using Firefly Feature Selection by Support Vector Machine (SVM). *International Journal of u-and e-Service, Science and Technology*, 8(5), 355-364.

Elberrichi, Z., & Aljohar, B. (2007). N-grams in Texts Categorization. *Scientific Journal of King Faisal University (Basic and Applied Sciences)*, 8(2), 1428H.

Fatima, S., & Srinivasu, B. (2017). Text Document categorization using support vector machine.

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., & Tserpes, K. (2012, June). Representation Models for Text Classification: a comparative analysis over three Web document types. *In Proceedings of the 2nd international conference on web intelligence, mining and semantics* (p. 13). ACM.

Hammami, M., Chahir, Y., & Chen, L. (2003, October). WebGuard: Web based adult content detection and filtering system. *In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on* (pp. 574-578). IEEE.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4), 784-796.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. *In ACM SIGIR Forum* (Vol. 51, No. 2, pp. 211-218). ACM.

Huang, C. C., Chuang, S. L., & Chien, L. F. (2004). Using a web-based categorization approach to generate thematic metadata from texts. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(3), 190-212.

Ishii, N., Murai, T., Yamada, T., & Bao,