International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I4P115

Volume 68 Issue 4, 81-86, April 2020
© 2020 Seventh Sense Research Group®

Review Article

A Systematic Review on Web Page Classification

Ajose-Ismail, B.M[1], Osanyin Q.A[2]

[1,2] *Department of Computer Science, School of Applied Science,The Federal Polytechnic Ilaro, Ogun State. Nigeria.*

Received Date: 06 March 2020
Revised Date: 23 April 2020
Accepted Date: 25 April 2020

Abstract *- With the increase in digital documents on the World Wide Web and an increase in the number of web pages 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 the 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 the 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 creating awareness of the need for an adequate document representation technique as this helps capture the semantics of document and also contribute to reducing the problem of high dimensionality.*

Keywords *- Bags of words model, Classification, Machine learning, Document representation, TF-IDF, Web Page classification, LDA, Word2Vec.*

I. INTRODUCTION

News and blogs webpages are today's most common sources for gathering information about current events. Information gotten from blogs and websites come in several categories in which users are only interested in certain topics within that category; for example, business, entertainment, sports or politics. Aggregating and categorizing information from these sources seems to be a daunting task as the volume of digital documents available online is growing exponentially as a result of increased usage of the internet [1]. Automated web categorization is the key technology for this task. Web page classification (WPC), also known as web page categorization, is the process of assigning a web page to one or more predefined category labels [2]. Web page classification problems can be divided into two categories: manual and automatic web page classification. Manual classification is a task that is performed by domain experts manually, and it looks impractical because it will take lots of human effort and time [3]. While automatic web page classification is a supervised machine learning problem where a set of documents is used to train the classifier, once training is done, it is used to classify web pages [4]. While the former is tedious and time-consuming, the latter saves a lot of manpower and material resources and time [5]. 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 [6], [7], [8]. Web pages are semi-structured documents formatted with HTML tags so that they may be rendered visually to users. Web documents also exist within a hypertext with connections within and outside the documents [6]. Several benefits can accrue from the accurate classification of documents into their respective categories, such as providing tools that help people to find, filter and analyze digital information 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 [9], the applications of WPC are as follows: Web directories provided by different search engines like Google, Yahoo amongst others, can be constructed, maintained or expanded using advanced WPC techniques [10], [11]. WPC is used to improve the quality of search results. When a user types in a particular keyword, the numbers of relevant results are increased through WPC [12]. A question-answer system uses WPC techniques to improve the quality of answers [13]. Web content filtering is another application of WPC [14]. Many WPC systems have been presented in literature over the years in which different perspectives have been taken to improve the performance of web classifiers [15]. Machine Learning (ML) algorithms such as Naïve Bayes, K-nearest neighbour, Decision tree, neural network, support vector machine and so on have been used previously by many researchers to achieve this task [16]. To achieve a high classification result of the WPC system, an excellent representation of textual data (Preprocessing) should contain as much information as possible from the original document [17]. Also, the accuracy of most classification algorithms depends on the quality and size of training data which is dependent on the document representation technique [3]. The general problem of web classification can be divided into three areas: document representation, classifier construction and classifier evaluation [18]. This paper provides an extensive study of the web page classification process with a thorough review of feature selection techniques used in document representation using existing literature. The remainder of this paper is organized as follows: Section 2 proceeds with an overview of the web page classification process highlights feature selection techniques used in the document representation phase of the WPC. In Section 3, related works on web page classification. Finally, the review is concluded in section 4, with also some future directions.

.

II. LITERATURE REVIEW

A. Web Page Classification Process

According to [16], the web page classification system is divided into several components, as shown in Figure 1 below. The stages of the web page classification process include: 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.

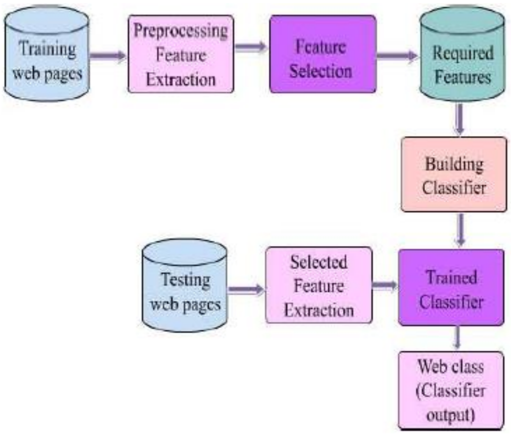

Fig. 1 Web page Classification Process
(Source: [19])

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, Javascript, Video Clips, Embedded objects, advertisement, Google ad-sense [2]. The extracted web contents are used in creating a corpus of labelled web pages, i.e. training web pages, which would be utilized by the classifier to build the learning system [6]. Already existing corpuses such as Reuters [3], [16], WebKb [4], [5], [9], Yahoo news dataset [20], sentiment Treebank [21], 20 Newsgroup dataset [17], IMDb dataset [22] can be utilized for this process or by creating a custom made corpus (using web crawlers) which can be used to automatically download web content [2].

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 [11]. The pre-processing stage can be further divided into Feature Extraction (FE) and Feature Selection (FS) [18]. FE process begins by extracting the raw content of the pages and discarding HTML tags and other WWW contents. Web page documents are characterized by high dimensionality. The first technique to reduce this high dimensionality is FE [1], [4]. Then FE continues by performing tokenization (breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens), stemming / lemmatization [16]. After feature extraction, the next step in the pre-processing stage is Feature Selection (FS) which involves constructing a vector space model of the document to improve the scalability, efficiency and accuracy of a text classifier. The main idea of FS is to select a subset of features from the original documents [19]. Also, with the inherent characteristics of web documents which are high dimensional datasets, FS is used to reduce the feature space and improve the efficiency and accuracy of classifiers. Feature selection approaches can be broadly classified as filter, wrapper, and embedded. The most generic of all the approaches is the filter approach, and it works irrespective of the data mining algorithm that is being used [3]. It typically employs measures like correlation, entropy, mutual information, and so forth, which analyzes general characteristics of the data to select an optimal feature set. However, it is to be noted that wrapper and embedded methods often outperform filters in real data scenarios [19]. In the embedded approach, feature selection is a part of the objective function of the algorithm itself. Similar examples can be seen in a decision tree, LASSO, LARS, 1-norm support-vector, and so forth. In contrast to the above approaches, which specifically select a subset of features, other techniques decompose the original higher-dimensional document-feature matrix into a lower-dimensional matrix, which effectively transforms the original feature space (determining the semantic relations between words, which defines the concept in the document) [6]. Also, they are deficient in revealing the inter-or intra-document statistical structure of the corpus [23]. Such methods include a bag of words model TF-IDF, Latent Semantic Indexing (LSI), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Indexing (PLSI), Word2Vec [17]. 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. DR is a very important step in web classification because irrelevant and redundant features often degrade the performance of the classification algorithms both in speed and classification accuracy and also its tendency to reduce overfitting [19]. Also, this stage has gained more attention recently than any other component of the WPC because effective dimension reduction makes the learning task more efficient and saves more storage space [24].

c) Obtaining the Required Features

The next stage after pre-processing is to gather the required feature set for classification, which is usually achieved by creating a matrix representation of the document vectors, which would be fed to the classifier [19].

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 algorithms have been used for building the model of the WPC system systems, such as KNN, Support Vector Machine (SVM), Artificial Neural Network (ANN), Deep Learning [15] and so on. After training the classifier, the model obtained is then used to automatically classify new web pages 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 depend on the training data set, i.e. the choice of the techniques used in the pre-processing stage have an overall effect on the quality of the classifier [25].

e) Evaluating the Classifier

An evaluation measure is used to measure the performance of a WPC classifier. For each category T_n, a confusion matrix can be constructed as shown in Figure 2 where ‘i’ denotes the number of true positive classifications, ‘j’ denotes the number of false-positive classifications, ‘k’ denotes the number of false-negative classifications and ‘l’ denotes the number of true negative classifications. For a perfect classifier, j and k would both be zero [26].

Table 1. Confusion Matrix

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

(Source: [26])

B. Feature Selection Techniques Used in WPC

According to the Google index, the volume of digital documents available online is over 130 trillion pages, and it's growing exponentially as a result of increased usage of the internet. Finding relevant and timely information from these documents is important for many applications [27]. The accuracy and generalization capabilities of the classifier in assigning a web page to its correct category is heavily dependent on the document representation [17]. 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 with one challenge or the other. Some of them are highlighted below:

a) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a model for document representation that is often used in information retrieval. It is a model that evaluates how important a word is to 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. 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 is an unordered collection of words (Turney & Pantel, 2010).

b) Latent Semantic Indexing (LSI)

LSI is a popular information retrieval method that uses the linear algebraic indexing method to produce low dimensional representations by word co-occurrence [28]. It utilizes the Singular Value Decomposition (SVD) algorithm on the sparse TF-IDF vector matrix to create a denser matrix that approximately models the original document. It composes frequencies of terms as a term-document matrix. The term-document matrix is a sparse matrix whose rows correspond to terms and whose columns correspond to documents. LSI was used to solve the synonym and polysemy problem of TF-IDF. However, a major drawback of LSI is that it does not capture multiple meanings of a word, and it does not respect word order. Also, LSA assumes that documents and features form a joint Gaussian model, while a Poisson distribution is typically observed, and the resulting dimensions might be difficult to interpret (Zhang, Yoshida & Tang, 2011).

c) Probabilistic Latent Semantic Indexing (PLSI)

To overcome some of the aforementioned problems with LSI, [29] introduced Probabilistic LSA (PLSA), which is a generative, graphical model enhancing latent semantic indexing (LSI) by a sounder probabilistic model. PLSI models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions. It uses EM Algorithm for its learning. PLSI is usually viewed as a more sound method as it provides a probabilistic interpretation, whereas LSI achieves factorization by using only mathematical foundations. The core of PLSA is a statistical model, which has been called the aspect model. Although PLSI had promising results, it suffers from two limitations: the number of parameters is linear in the number of documents, and it is not possible to make inferences for unseen data [23].

d) N-GRAM MODEL

The N-gram model is one of the most widely used models for feature representation. It assumes that the probability of a given the word is only conditional on its preceding n-1 word, where n could be 1 (the unigram model), 2 (the bigram model), 3 (the trigram model), or any whole number. This approach converts a collection of text documents into feature vectors by recording the n-gram frequency counts and uses the vectors as input to classifiers. According to (Elberrichi & Aljohar, 2007), some of the major strengths of N-GRAM are no need to perform word segmentation, automatic capture of the roots of the most frequent words, independence towards the document language, tolerances with the spelling mistakes and the deformations. In addition, no dictionary or language-specific techniques are needed. Also, N-GRAM suffers from data sparcity and high dimensionality [30].

e) *Latent Dirichlet Analysis (LDA)*

LDA is a probabilistic topic model that generates topics based on word occurrences from a corpus or set of documents [31]. 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 particularly useful for finding reasonably accurate mixtures of topics within a given document set. LDA is an unsupervised language model that transforms words from a bag of words counts into a continuous representative matrix. Also, LDA uses an unsupervised learning function that depends on words in the corpus, which will determine the matching degree and thus will suffer from vocabulary mismatching problems [32].

f) **Word2Vec**

Word2Vec is a Recurrent Neural Network based implementation that can learn word embeddings. The 2 main architectures are CBOW (Continuous Bag-of-word) and Skip-gram (Continuous Skip-gram Model). CBOW 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 w (t) is a column in the Matrix W. The CBOW model predicts a word w (t) utilizing the context w (t − n)... w (t − 1), w (t + 1)..., w (t + n), while the Skip-gram model predicts each word in the context utilizing the word w (t). The Word2Vec framework aims at predicting the context of a word or word based on their context. The word embeddings are learned through maximizing the objective function. With these word embedding's it can capture distributed representations of text to capture similarities among concepts [33] 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, Ma & Zhang, 2016).

III. RELATED WORKS ON WEB PAGE CLASSIFICATION

In the works of [1], they proposed a method to accurately and automatically classify web pages into different categories viz three phases: feature extraction, information learning and classification. In the methodology adopted, a term-document matrix is created using tf-idf, then the terms are used to extract object-based features. A decision tree algorithm is then used to extract rules from the features extracted. The web pages are then classified using optimal firefly algorithm based Naive Bayes Classifier (FA-NBC) using the rules extracted. The proposed method was applied to WebKB datasets. Experimental results show 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 [17].

[34] proposed a method to analyze and categorize e-commerce websites automatically. In their methodology, e-commerce websites were crawled, text preprocessing and the terms of the document were derived using tf-idf. The proposed method was applied to 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 show that the produced method outperforms pure TF-IDF. Also, the results show 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 [25].

In the works of [17], they proposed the use of a hybrid strategy that consists 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 were applied to 20 Newsgroup data using an SVM classifier. Results obtained show that their proposed methods outperform earlier methods such as TF-IDF+ SVM, Word2Vec + SVM, LDA + SVM. One of the major drawbacks of their method is that improper calibration of the LDA parameters (e.g. a number of topics, hyper-parameters) could potentially lead to sub-optimal results as most of the parameters for the LDA are imported from the natural language community.

[15] proposed the categorization of web News using word2vec and deep learning. Earlier methods of automatic classification used supervised and unsupervised algorithms such as SVM, Naive Bayes and clustering, respectively but are marred with several issues. The former can only handle supervised data, but actual texts on the web are not supervised data. The latter defines the category automatically. Also, another problem with web classification is transforming the text into constant dimensions. In their work, Word2Vec was used to train the news corpus into vectors. After obtaining the word vectors, pre-training of the vectors using autoencoder and training of the dataset using deep learning framework. Experimental tests were carried out on a web news site (Yahoo) containing 1,728,942 records, and results obtained show that deep learning produced a better result than Naive Bayes but performed badly on training time.

IV. OBSERVATIONS

Representation of the input data (DR) is a crucial issue in web page classification and text classification systems at large. 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 [26]. Also, from extant literature, SVM, KNN, and Naïve Bayes are some of the most widely used ML algorithms for text classification [1], [5], [24]. In the work of [3], they decided to investigate this issue and compared SVM, KNN and Naïve Bayes on text classification tasks. Results obtained show 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 scale up to the performance of SVM. In light of this, there is a 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 approach has been shown to perform better (obtain the semantic features) by harnessing the strength of the individual technique in the arrangement Word2Vec and LDA [17] or Word2Vec and TF-Idf [24]. Also, proper calibration of the parameters of LDA with a search algorithm would produce better latent topics across words in a document [32].

V. CONCLUSION

In this paper, we gave an overview of the 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 into with a view to identifying 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 the document representation phase is one of the areas that are receiving interest from researchers. The most currently used methods of document representation are the Vector Space Model (VSM), Probabilistic Topic Model and Statistical Language Models and Neural network language models [25]. The chosen document representation technique has a direct impact on the classification results as it captures the semantics of the document and also contributes to reducing the problem of high dimensionality. Combining different DR techniques is a new area of research because each technique performs differently depending on the dataset. Future work in WPC should focus on improving the semantic relationship of a web document by hybridizing the difference DR technique which will inherently improve the classification result. Also, ontology-based techniques can be used to capture the real semantics of web documents.

REFERENCES

[1] Raj, A. J., Francis, F. S., & Benadit, P. J., Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC, Computer Science and Engineering, 6(1) (2016) 7-13.

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

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

[4] Shibu, S., Vishwakarma, A., & Bhargava, N., A combination approach for web page Classification using Page Rank and Feature Selection Technique, International Journal of Computer Theory and Engineering, 2(6) (2010) 897.

[5] Dixit, S., & Gupta, R. K., 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) (2015) 355-364.

[6] Qi, X., & Davison, B. D., Web page classification: Features and algorithms. ACM computing surveys (CSUR), 41(2) (2009) 1-31.

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

[8] AbdulHussien, A. A., Comparison of Machine Learning Algorithms to Classify Web Pages. International Journal of Advanced Computer Science and Applications (IJACSA), 8(11) (2017).

[9] Mangai, J. A., Kothari, D. D., & Kumar, V. S., A Novel Approach for Automatic Web Page Classification using Feature Intervals, International Journal of Computer Science Issues (IJCSI), 9(5) (2012).

[10] Huang, C. C., Chuang, S. L., & Chien, L. F., Using A Web-Based Categorization Approach To Generate Thematic Metadata From Texts. ACM Transactions On Asian Language Information Processing (TALIP), 3(3) (2004) 190-212.

[11] Mangai, J. A., Kumar, V. S., & Balamurugan, S. A., A Novel Feature Selection Framework For Automatic Web Page Classification. International Journal Of Automation And Computing, 9(4) (2012) 442-448.

[12] Haveliwala, T. H., Topic-Sensitive Pagerank: A Context-Sensitive Ranking Algorithm For Web Search. IEEE Transactions On Knowledge And Data Engineering, 15(4) (2003) 784-796.

[13] Cui, H., Kan, M. Y., Chua, T. S., & Xiao, J., A Comparative Study on Sentence Retrieval for Definitional Question Answering, in SIGIR Workshop on Information Retrieval For Question Answering (IR4QA), (2004) 383-390.

[14] Hammami, M., Chahir, Y., & Chen, L., Webguard: Web-Based Adult Content Detection and Filtering System, in Web Intelligence, WI,Proceedings. IEEE/WIC International Conference On, (2003) 574-578. IEEE.

[15] Kato, R., & Goto, H., Categorization of Web News Documents Using Word2vec And Deep Learning, In Proceedings of The 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia, (2016).

[16] Fatima, S., & Srinivasu, B., Text Document Categorization Using Support Vector Machine, (2017).

[17] Wang, Z., Ma, L., & Zhang, Y., A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec, In Data Science in Cyberspace (DSC), IEEE International Conference On, (2016) 98-103. IEEE.

[18] Khan, A., Baharudin, B., Lee, L. H., & Khan, K., A Review Of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1) (2010) 4-20.

[19] 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) (2010) 112-122.

[20] Yin, D., Hu, Y., Tang, J., Daly, T., Zhou, M., Ouyang, H., & Langlois, J. M., Ranking Relevance In Yahoo Search, In Proceedings of The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016) 323-332. ACM.

[21] Socher, Richard, Perelygin, Alex, Wu, Jean Y., Chuang, Jason, Manning, Christopher D., Ng, Andrew Y., and Potts, Christopher., Recursive Deep Models for Semantic Compositionality over A Sentiment Treebank, on Conference on Empirical Methods In Natural Language Processing, (2013).

[22] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C., Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1 (2011) 142-150. Association for Computational Linguistics.

[23] Biro, I., Benczur, A., Szabo, J., & Maguitman, A., A Comparative Analysis of Latent Variable Models For Web Page Classification, In Latin American Web Conference, LA-WEB'08, (2008) 23-28. IEEE.

[24] Lilleberg, J., Zhu, Y., & Zhang, Y., Support Vector Machines And Word2vec for Text Classification With Semantic Features, In Cognitive Informatics & Cognitive Computing (ICCI* CC),IEEE 14th International Conference On, (2015) 136-140. IEEE.

[25] Singh, K. N., Devi, H. M., & Mahanta, A. K., Document Representation Techniques And Their Effect On The Document Clustering And Classification: A Review. International Journal Of Advanced Research In Computer Science, 8(5) (2017).

[26] Jindal, R., Malhotra, R., & Jain, A., Techniques for Text Classification: Literature Review and Current Trends. Webology, 12(2) (2015) 1.

[27] Azam, N., & Yao, J., Comparison of Term Frequency and Document Frequency-Based Feature Selection Metrics In Text Categorization, Expert Systems with Applications, 39(5) (2012) 4760-4768.

.

[28] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R., Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6) (1990) 391.

[29] Hofmann, T., Probabilistic Latent Semantic Indexing. In ACM SIGIR Forum, 51(2) (1999) 211-218. ACM

[30] Le, Q., & Mikolov, T., Distributed Representations of Sentences and Documents, in International Conference on Machine Learning, (2014) 1188-1196.

[31] Blei, D. M., Ng, A. Y., & Jordan, M. I., Latent Dirichlet Allocation, Journal Of Machine Learning Research, 3 (2003) 993-1022.

[32] Dit, B., Panichella, A., Moritz, E., Oliveto, R., Di Penta, M., Poshyvanyk, D., & De Lucia, A., Configuring Topic Models for Software Engineering Tasks in Trace Lab, In Traceability in Emerging Forms of Software Engineering (TEFSE),International Workshop on, (2013) 105-109. IEEE.

[33] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations In Vector Space. Arxiv Preprint Arxiv:1301.3781.

[34] Moiseev, G., Classification of E-Commerce Websites By Product Categories, In AIST (Supplement), (2016) 237-247.

[35] Turney, P. D., & Pantel, P., from Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research, 37 (2010) 141-188