



SDI Review Form 1.6

Journal Name:	<u>British Journal of Applied Science & Technology</u>
Manuscript Number:	Ms_BJAST_30376
Title of the Manuscript:	A Design of an Automatic Web Page Classification System
Type of the Article	Original Research Article

General guideline for Peer Review process:

This journal's peer review policy states that **NO** manuscript should be rejected only on the basis of '**lack of Novelty**', provided the manuscript is scientifically robust and technically sound.

To know the complete guideline for Peer Review process, reviewers are requested to visit this link:

(<http://www.sciencedomain.org/page.php?id=sdi-general-editorial-policy#Peer-Review-Guideline>)



SDI Review Form 1.6

PART 1: Review Comments

	Reviewer's comment	Author's comment (if agreed with reviewer, correct the manuscript and highlight that part in the manuscript. It is mandatory that authors should write his/her feedback here)
Compulsory REVISION comments	General remark for the joined words in the whole manuscript. (Example: "This paper proposes a hybrid Web page classifier based on the classical Naive Bayesian, K-Nearest Neighbor and Support Vector Machine.") Section 2. Related work – needs major revision Needs introductory for the problem that you are reviewing. New publications have to be considered for Automatic Web Page Classification. Peng and Choi have proposed single-path search technique to reduce the search complexity and increases the accuracy for text classification of Web pages [1]. 1. Xiaogang Peng and B. Choi, "Automatic web page classification in a dynamic and hierarchical way," Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, 2002, pp. 386-393. doi: 10.1109/ICDM.2002.1183930 Also, neural network classifiers [2] should be mentioned. 2. Thangairulappan, K. and Kanagavel, A. (2016) Improved Term Weighting Technique for Automatic Web Page Classification. Journal of Intelligent Learning Systems and Applications, 8, 63-76. doi: 10.4236/jilsa.2016.84006. Section 4. Elaborate why did you choose Bayes, SVM and KNN classifiers? Section 5. It looks like web page classification methodology is missing with clear description, such as "In text-based approaches, first a database of keywords in a category is prepared as follows. The frequency of the occurrence of words, phrases etc in a category is computed from an existing corpus (a large amount of text). The commonly occurring words (called stop words) are removed from this list. The remaining words are the keywords for that particular category and can be used for classification. To classify a document, all the stop words are removed and the remaining keywords/phrases are represented in the form of a feature vector. This document is then classified into an appropriate category using the classification	



SDI Review Form 1.6

	algorithm." source: Xiaoguang Qi and Brian D. Davison. 2009. Web page classification: Features and algorithms. ACM Comput. Surv. 41, 2, Article 12 (February 2009). DOI=http://dx.doi.org/10.1145/1459352.1459357 Section 5. It is confusing 3-rd sentence "In the Classification phase, we use the Naive Bayes algorithm to classify the Web sites." And in Section 5.3. Improving phase you are mentioning "In our implementation, Naïve Bayesian algorithm is used in the feature selection process and then the support vector machine is used in the classification process." In Section 5.2. Testing phase you mention 5 categories (.Earrings, Mergers, Money, Grain and Crude Oil). The first category Earrings is not consistent with the description given in 5.4. and with the data given in tables 2,3,4,5,6,7. Update the link in reference [25] for the dataset. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html	
--	---	--



SDI Review Form 1.6

<u>Minor</u> REVISION comments	In Section 5.2. Classification Phase Phase 4- Extract the most interesting features from the list of features. What is the criteria and how do you choose the most interesting features? In Section 5.1 in the 3-rd step - Remove stop words and stemming, I would suggest to add lemmization as fine tuning process. By definition the Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. source: http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html Section 6.3.4. Experiment 4. It is not mentioned in which order NB, KNN and SVM classifiers are applied, as in Section 6.3.2.	
<u>Optional/General</u> comments	Section 2. Add more relevant publications. J. Eberius, K. Braunschweig, M. Hentsch, M. Thiele, A. Ahmadov and W. Lehner, "Building the Dresden Web Table Corpus: A Classification Approach," 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), Limassol, 2015, pp. 41-50. doi: 10.1109/BDC.2015.30 Gao Kening, Yang Leiming, Zhang Bin, Chai Qiaozi and Ma Anxiang, "Automatic classification of Web information based on site structure," 2005 International Conference on Cyberworlds (CW'05), Singapore, 2005, pp. 7 pp.-558. doi: 10.1109/CW.2005.24 Section 4. Introduction for text based categorization is useful and emphasize the usefulness of supervised classifiers for your approach.	

Reviewer Details:

Name:	<i>Aleksandar Karadimce</i>
Department, University & Country	<i>Computer Science and Engineering, University of Information Science and Technology "St. Paul the Apostle", Ohrid, R. Macedonia</i>