JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Comparative analysis of Machine learning techniques for Webpage classification

[1] Ratan Kumar Sajja ,[2]Siva Jyothi Barla,

[1]Associate Professor ,[2]DAssistant Professor,
[1]Computer Science &Engineering , [2] Computer Science & Engineering,
[1]Anil Neerukonda Institute of Technology & Sciences, Vizag, Inida,
[2]Anil Neerukonda Institute of Technology & Sciences, Vizag, Inida

Abstract : The World Wide Web is the largest informational archive that is accessible from anywhere in the world. It is becoming more and more crucial to classify websites based on their content as a result of the development in both the number of websites and online visitors. A content-based website classification is required, according the current trend, as recognizing a website only based on its URL is insufficient to obtain the desired results. Machine learning techniques can be used to find the required data on the content of web pages. In this work, we analyzed the content of web pages using a range of machine learning (ML) methods, such as SVM, Random Forest, Naive Bayes, Logistic Regression, Gradient Boosting, and AdaBoost. With a score of 0.982, AdaBoost outperformed the other algorithms in terms of Classification Accuracy, Precision, F1, and Recall.

IndexTerms **- Web Content Mining, Web Page Classification, Machine-learning algorithms**

I. INTRODUCTION

The Word Wide Web is a large repository of information which can be shared and accessed throughout the world. It is necessary to find out the suitable web page for the user request. The Web Content Mining(WCM) solves this and provides the way to access the content of web in a relevant way. WCM concerned with retrieving, classifying and discovering knowledge from web page content.

Despite the WCM techniques currently in use, a keyword search presents users with a massive amount of suggestions. The web page classification is necessary to address this problem. Web content filtering, ontology annotation, assisted web contextual advertising and knowledge base construction, building, maintaining, or expanding web directories (web hierarchies), assisting question answering systems to improve the quality of search results, developing effective focused crawlers or vertical (domain-specific) search engines, and improving search results are just a few of the many applications of web page classification.

In our work, web site classification is done using labels from pre-defined categories. Afterward, based on its content, the website is put into one or more categories. Adult, Business/Corporate, Computer & Technology, E-commerce, Food, Forums, Games, Health-care, Law and Government, News, Photography, Social Networking and Messaging, Sports, Streaming, and Travel are the categories into which the websites are divided. Websites were assigned to one or more of these categories by applying the ML algorithms. SVM, Random Forest, Naïve Bayes, Logistic Regression, Gradient Boosting, and AdaBoost algorithms were used in our research to classify the webpage to one or more of these categories depending on its content. Finally, we draw the conclusion that, when compared to other algorithms, the AdaBoost method produces results that are more accurate.

The remaining sections of the document are included below. The Literature Review is described in Section 2, the Proposed Method is explained in Section 3, the Results are shown in Section 4, and the Conclusion is discussed in Section 5.

II. LITERATURE REVIEW

Web mining is proposed by [1], he showed the way how the data mining techniques can be applied to retrieve the knowledge from web content. According to [2] Retrieving proper web page based on the content is a challenging task due to lack of structure in web resources. Authors of [3][4], explained about web content mining and semantic web mining. According to them the challenges that need to addressed are getting the relevant content for the user request. Authors of [5], discussed about Web Content Mining (WCM), Web Structure Mining (WSM), Web Usage Mining (WUM) and how they builds up the web mining.

Authors of [6], used Bag of Words (BOW) in Content Based Image Retrieval System for image annotation. Along with the usage of SVM this approach, produce F-measure result as 0.809-0.878.

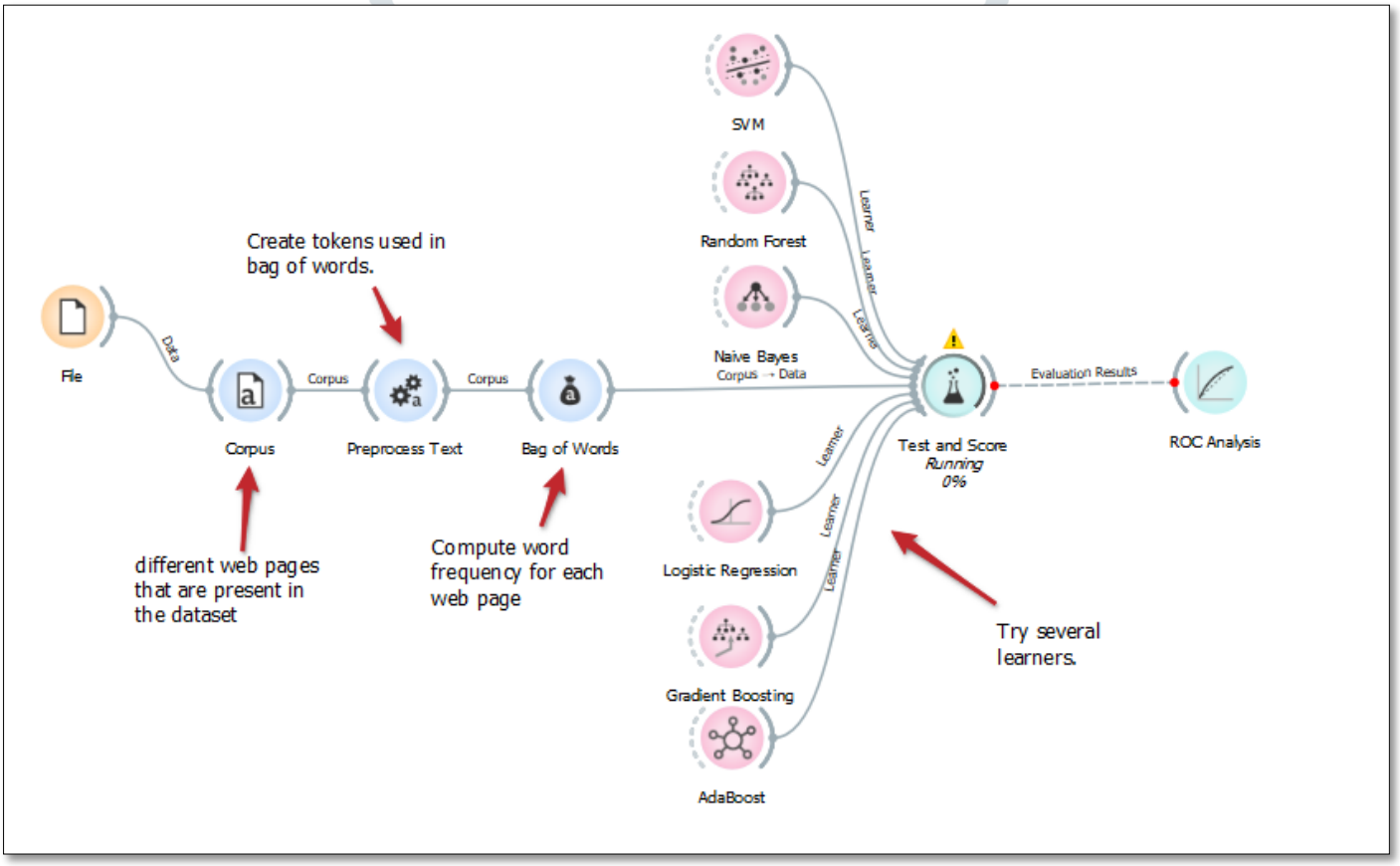
Authors of [7], used SVM and BOW to identify the sentiment analysis in a document at sentence level. Using these techniques, they are able to identify whether the text is sentimental and which class of emotion it belongs with an accuracy of 81.44-85.36. Authors of [8], produces a hierarchical web page classification using SVM model by integrating the current web page with its neighboring page vial Topic model which gives an accuracy equals to 90.33% and the F1 measure 90.14%. Authors of [9], presented A Web Page Classification Algorithm and Its application in E-government System, for classification of webpage, they combined the Unsupervised Clustering method with Support Vector Machine. They concluded that this approach gives a good result and it can be applied to e-commerce in future.

Authors of [10], used an approach to categorize the webpage using Map Reduce Programming Model. In this approach the crawled web pages submitted as input to the MRPM model to find the appropriate category based on the webpage content. This model produces the F-measure result as 0.720-0.910.Authors of [11], Webpage classification had done by applying a deep leaning based system. In this paper they used a RNN architecture to test. Using this model the success rate of the classification is approximately 85%. In [12], Authors presented Ensemble approach for web page classification, used a pre-trained model BERT for classifying the web pages into various categories. They included Nonlinear layer, CNN, DRIMN with BERT model and the results are describing that BERT base DRIMN is giving low error rate and high accuracy. In[13], authors using Machine Learning for web page classification in search engine optimization, classifies the unknown web pages into three predefined classes. The results show that the machine learning classifier accuracy ranges from 54.59% to 69.67%.

The available literature clearly shows that different machine learning techniques are applied for web page classification and there is a need of finding better models, which can give high accuracy.

III PROPOSED METHOD:

We considered each web page present in the dataset as a document, applied Bag of words to convert the text, and prepared vector representation of the web page to apply machine-learning algorithms as shown figure 3.1



3.1 Proposed method

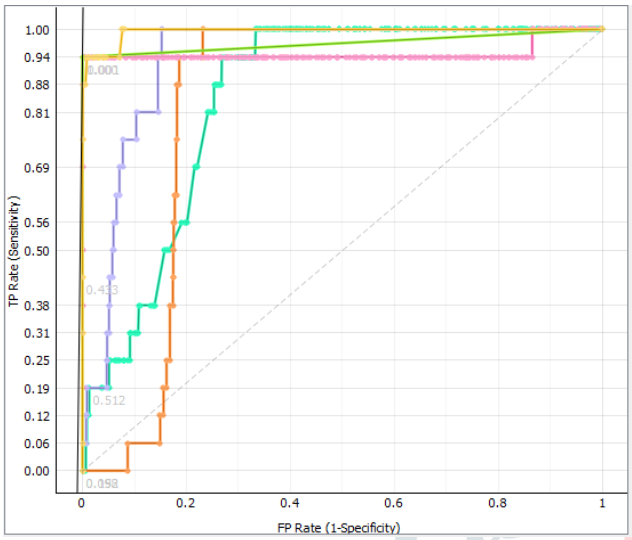
IV RESULTS

We took the publically available dataset [14] from kaggle website and applied SVM, Random Forest, Naive Bayes, Logistic Regression, Gradient Boosting, and AdaBoost machine learning algorithms with 5 fold cross validation. The standard metrics like Classification Accuracy (CA), F1 score ,Precision and Recall are taken into consideration to understand the different machine learning algorithms and the results shown in the table 4.1. The dataset [14] has 1425 samples and with 2.4% of missing data. AdaBoost Algorithm clearly outperformed all the other algorithms.

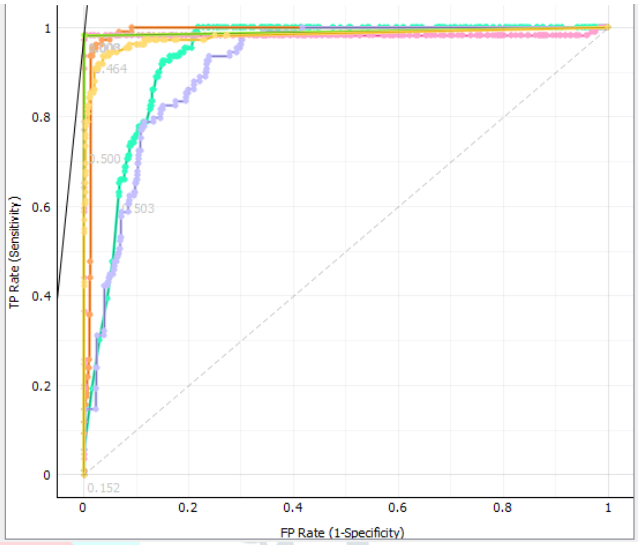
Table 4.1: Machine Learning Algorithms performance

Model	CA	F1	Precision	Recall
SVM	0.081	0.012	0.007	0.081
Random Forest	0.812	0.809	0.815	0.812
Naïve Bayes	0.522	0.502	0.563	0.522
Logistic Regression	0.836	0.826	0.822	0.836
Gradient Boosting	0.978	0.977	0.978	0.978
AdaBoost	0.982	0.982	0.982	0.982

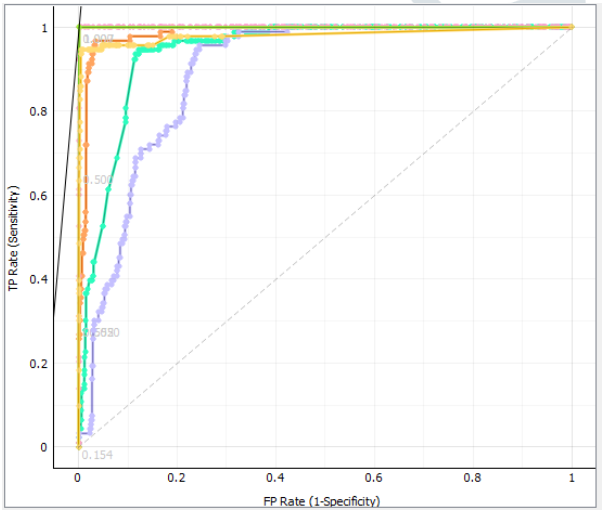
The ROC curves are presented here show sensitivity and specificity for each target class.



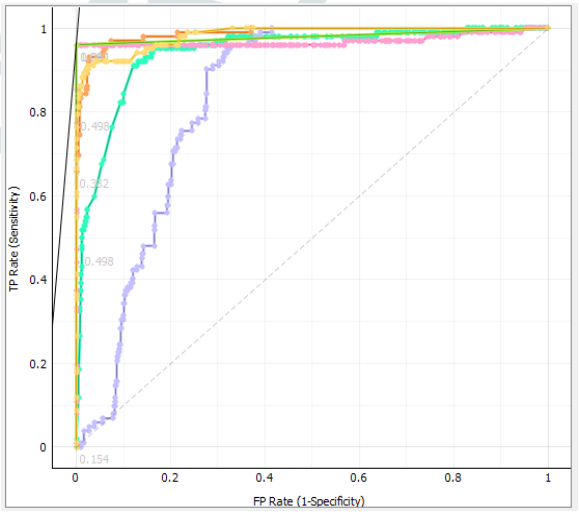
ROC curve for Class Label: Adult



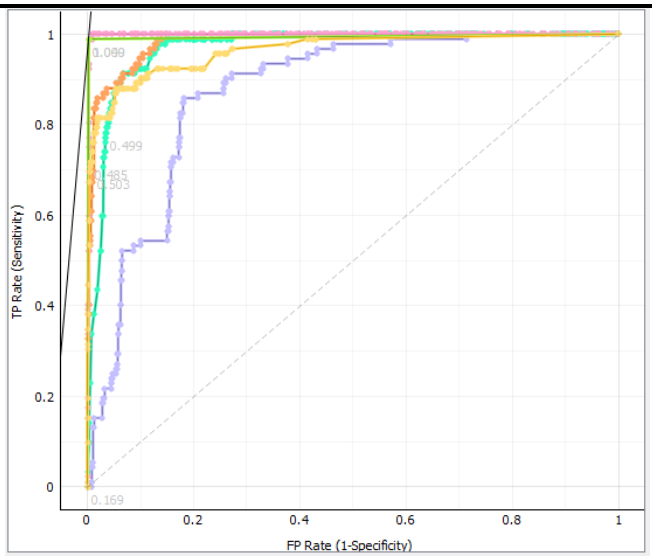
ROC curve for Class Label: Business/Corporate



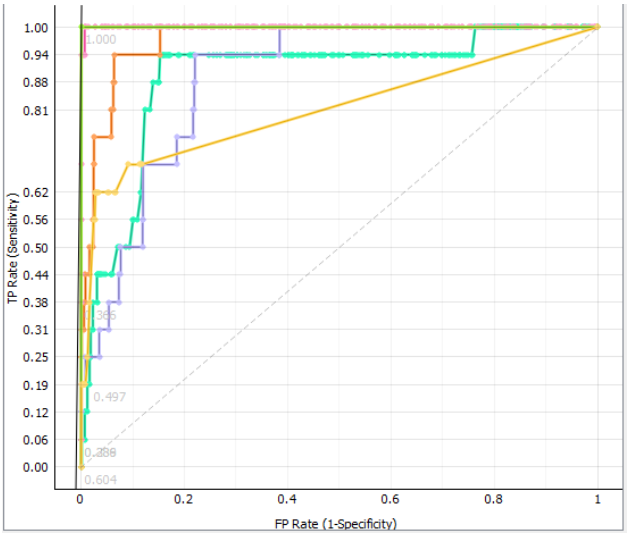
ROC curve for Class Label: Computers and Technology



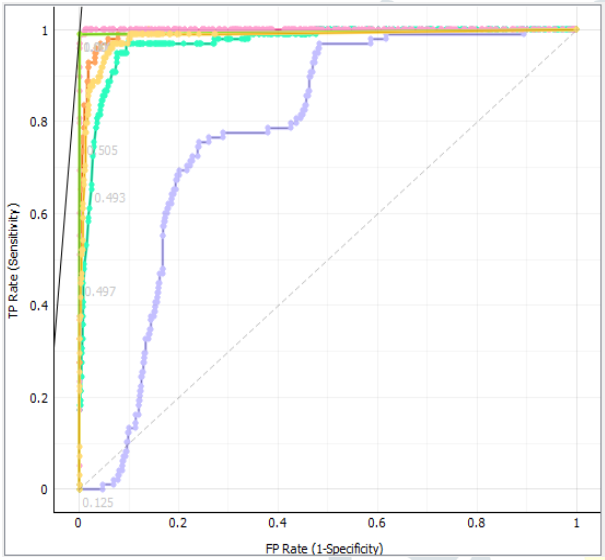
ROC curve for Class Label: E-Commerce



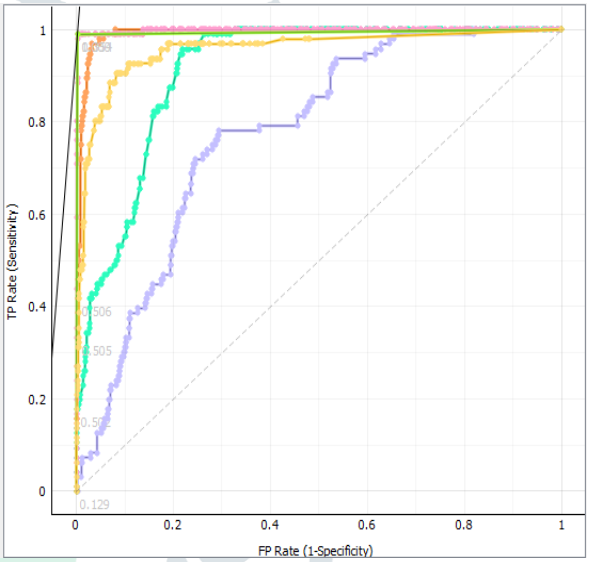
ROC curve for Class Label: Food



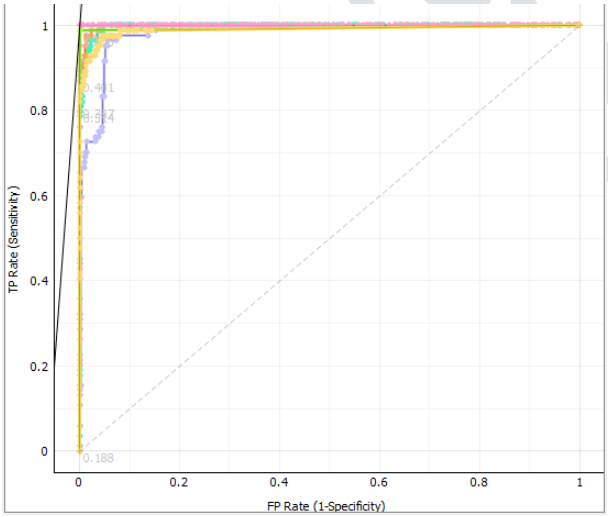
ROC curve for Class Label: Forums



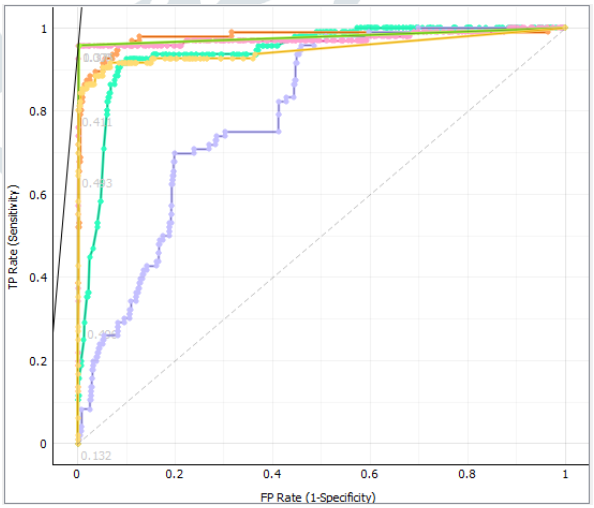
ROC curve for Class Label: Games



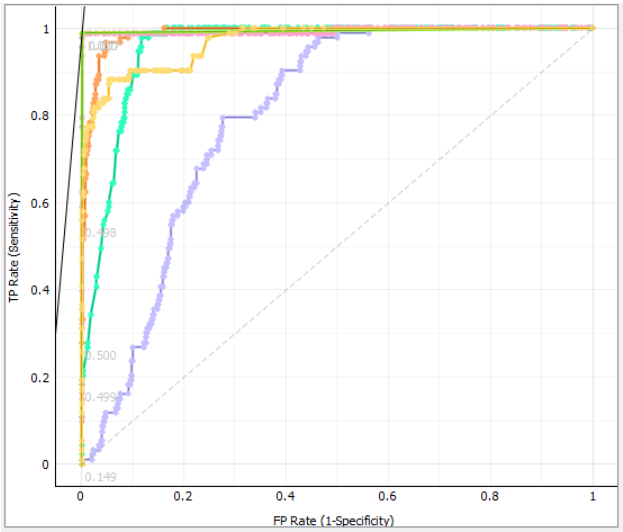
ROC curve for Class Label: Health-Care



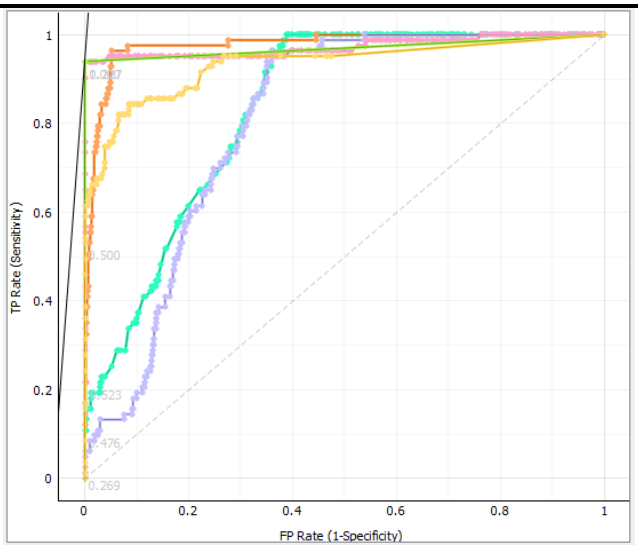
ROC curve for Class Label: Law and Government



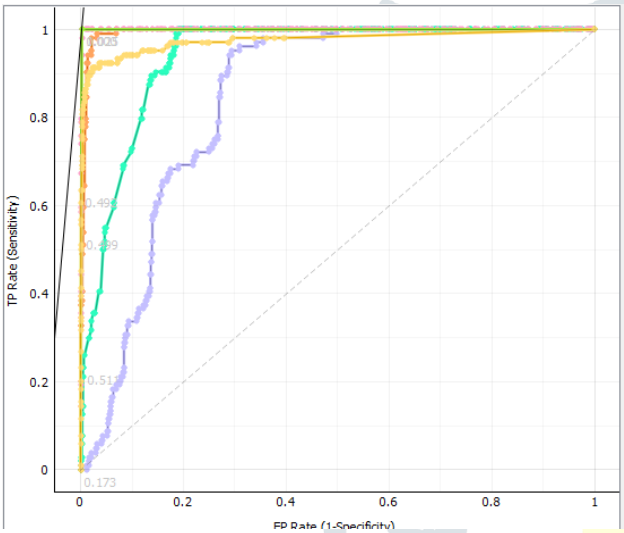
ROC curve for Class Label: News



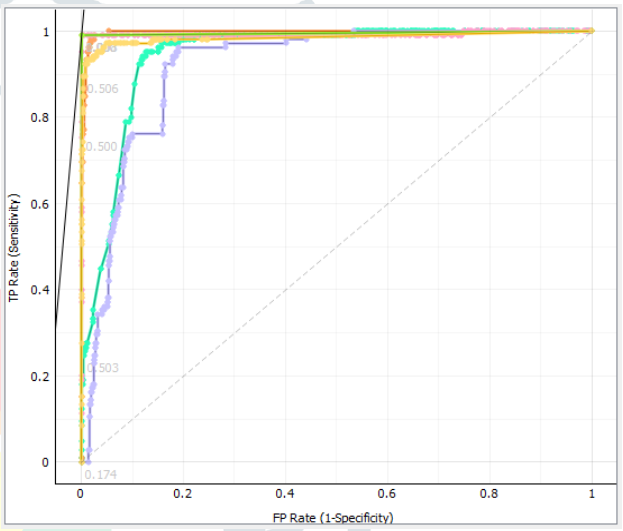
ROC curve for Class Label: Photography



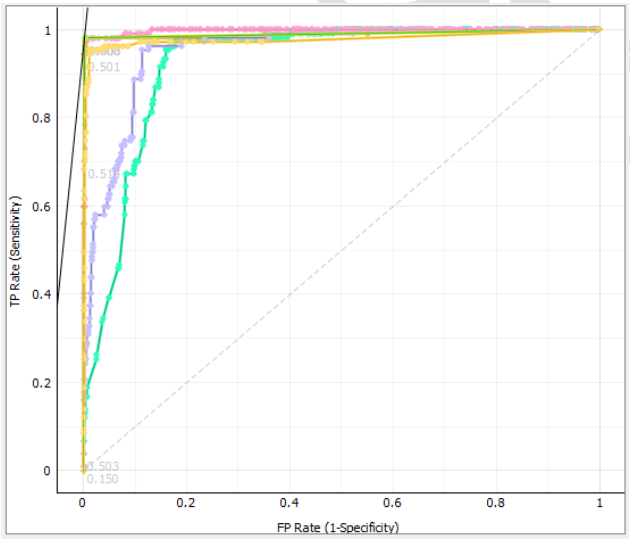
ROC curve for Class Label: Social Networking and Messaging



ROC curve for Class Label: Sports



ROC curve for Class Label: Streaming



ROC curve for Class Label: Travel

V CONCLUSION

Web page classification is becoming one of the major research areas in the contemporary world with rise of World Wide Web. In the present work, bag of words method is applied to convert to web page content into vector representation so that different machine learning techniques can be applied. We found that AdaBoost algorithm is able to produce high accuracy. We are planning to apply document embedding techniques methods in place of word of bags and apply different machine learning algorithms as our future work.

References

1. Etzioni, O (1996). The World Wide Web: Quagmire or gold mine? Communications of the ACM, 39(11), 65–68.
2. Liu, B (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd edn., pp. 55–235. Secaucus, USA: Springer-Verlag, New York, Inc
3. Kolari, P and A Joshi (2004). Web mining: Research and practice. IEEE J Computing in Science and Engineering, 6(4), 49–53
4. Yong-gui, W and J Zhen (2010). Research on semantic web mining. In CDA 2010: Proceedings of International Conference, China, Vol. 1, pp. 67–70, 25–27 June
5. Bin, W and L Zhijing (2003). Web mining research. In CIMA 2003: Proceedings of the 5th International Conference, IEEE Computer Society, China, pp. 165– 170, 27–30 September.
6. Xu, Z, I King and M Lyu (2007). Web page classi¯cation with heterogeneous data fusion. In WWW 2007: Proceedings of the 16th International Conference, Alberta, Canada, pp. 1171–1172, 8–12 May.
7. Chen, CL, HM Lee and Y Chang (2009). Two novel feature selection approaches for web page classi¯cation. Elsevier ESA, 36, 260–272.
8. Sriurai, W, P Meesad and C Haruechaiyasak (2010). Hierarchical web page classi¯cation based on a topic model and neighboring pages integration. International Journal of Computer Science and Information Security, 7(2), 166–173.
9. Boyi Xu, Jing Wang, Hongming Cai (2010) A Web Page Classification Algorithm and Its Application in E-government System. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)
10. Malarvizhi, P and VR Pujeri (2012). Distributed approach to web page categorization using map-reduce programming model. International Journal of Engineering and Technology, 3(6), 373–38
11. Ebubekir BUBER, Banu DIRI, (2019) Web Page Classification Using RNN, 8th International Congress of Information and Communication Technology (ICICT-2019)
12. Amit Gupta, Rajesh Bhatia, (2021). Ensemble approach for web page classification, Multimedia Tools and Applications (2021) 80:25219–25240
13. Goran Matosevic, Jasminka Dobsa and Dunja Mladenic, (2021), Future Internet 2021, 13, 9: 2-20
14. https://www.kaggle.com/hetulmehta/website-classification