EasyChair Preprint

№ 2188

The Effect of Bloom's Taxonomy on Random Forest Classifier for cognitive level identification of eLearning content

Benny Thomas and J Chandra

December 18, 2019

The Effect of Bloom's Taxonomy on Random Forest Classifier for cognitive level identification of eLearning content.

Mr.Benny Thomas
Department of Computer Science,
Christ (Deemed to be University), Bangalore, India
Benny.thomas@christuniversity.in

Dr. Chandra J
Department of Computer Science,
Christ (Deemed to be University), Bangalore, India
Chandra.j@christuniversity.in

Abstract. With the advancement in internet, the efficiency of e-learning increased and currently e-learning is one of the primary method of learning for most learners after the regular academics studies. The knowledge delivery through e-learning web sites increased exponentially over the years because of the advancement in internet and e-learning technologies. The learner can find many website with lots of information on the relevant domain. However learners often found it difficult to figure out the right leaning content from the humongous availability of e-content. In the proposed work an intelligent framework is developed to address this issue. The framework recommend the right learning content to a user from the e-learning web sites with the knowledge level of the user. The e-contents available in web sites were divided in to three cognitive levels such as beginner, intermediate and advanced level. The current work uses Blooms Taxonomy verbs and its synonyms to improve the accuracy of the classifier used in the framework.

Keywords—eLearning; Blooms Taxnomy; Random Forest Classifier ; Machine Learning; POS tagging; Decision trees;Cognitive level.

I. INTRODUCTION

E-learning is a prevalent method of learning with the help of internet and other e-leaning technologies which bridges the physical gap between leaner and the teacher. Nowadays E-leaning become more popular because of the availability of good e-leaning technologies and world class e-leaning web sites free of cost through internet. It is one of the primary mode of learning for most of the entrepreneurs and working professional. E-learning gives us the flexibility and choice to learn from anywhere and at any time. Because of its potentiality and broader usage the e-learning web sites increased tremendously over the years. Finding multiple e-learning web sites needed for a domain is easy these days. However most learners get overwhelmed by the enormity of content availability and find it difficult to figure out the right learning content. Contents w so tilted that the learner finds it too difficult to get the suitable learning material. Users spent lot of time to understand the content needed end up learning nothing to improve the knowledge.

The current work proposed an intelligent solution to address this problem. An intelligent e-content recommendation system based on the domain knowledge of the user is developed to provide the right learning content to the user from the free e-learning web sites. In the current work, Blooms Verbs and its synonyms were added to the extracted feature set as additional feature to improve the performance of the classification algorithm.

Bloom taxonomy divides a learning content in to different cognitive levels based on the difficulty level of the content[1]. To do this a set of Blooms verbs were defined which helps to identify the difficulty level of the content. The Random Forrest Classifier is used after comparing the performance of many text classification algorithms for Machine Learning.

The remaining part of the paper is organized as follows, Section II Literature review, Section III.Methodology, Section IV Implementation Result and Discussions and Section VI Conclusion.

II. LITERATURE REVIEW

Number of studies have conducted to recommend the right eLearning content to learners through various methods. The Blooms taxonomy is used in learning content classification by earlier researchers. Fatema Nafa et.al., used text analysis for automatic content classification with the help of Naive Base Classifier which identifies the Blooms Taxonomy levels in the text paragraph using rules in the training set. The text is split in to paragraph and the verb in between the noun is estimated. The validity of the verb is estimated using alpha theroshold[2]. Ursula Fuller et.al., proposed a computer science based learning taxonomy specially for computer science domain as the computer science domain is not well captured by the existing taxonomies. The work identified that the Blooms taxonomy dominate the computer science assessment design[3]. Amal Babour et.al propose a graph tringluarity based method for classifying the knowledge unites in textual graph that can identify the Blooms Taxonomy levels. A verb based relation extraction algorithm is used to extract relation between text and concept[4]. Fatema and Javed Khan proposed method to improve the quality of educational contents based on cognitive theory using Blooms Taxonomy as any knowledge domain can be learned and taught in multiple cognitive levels. The Blooms taxonomy levels between various relations and concepts were automatically extracted in this work. The method used verbs to find out Blooms taxonomy relationship domain knowledge[5]. Anwar Ali Yahya et.al., proposed a method to understand the cognitive levels of class room questions using machine learning. The questions were automatically classified to different cognitive levels identified by Blooms Taxnomy. The dataset used is the question papers collected and classified according to the cognitive levels[6]. Kyoung Mi Yang et.al., proposed a method to construct learning path through e-learning using Item response theory which refers to Blooms Taxonomy cognitive levels. It construct a discrete form of knowledge to be learned by high school and secondary school students[7] Numerous studies have conducted a study to recommend the best e-learning content to the learner using different text classification methods. Atorn Nuntiyagul el.al., used Patterned keywords and Phrase with support vector machine algorithms to classify the items. The approach uses the text classification techniques in machine learning and information retrieval[8]. G. Desai et.al., proposed that the Naiye Bayse method as one of the best method for document classification as it gives good results. The methodology used is random sampling of the labeled categories of text. The disadvantage of this method is that it does not consider the morphological structure of the terms used in the text[9]. Sankar Perumal et.al., proposed a new content recommendation which delivers best contents by refining the final frequent item patterns obtained from frequent pattern mining technique and then classifying the contents using fuzzy logic into three levels by generating frequent item pattern. It has higher efficiency compared to the other similar methods. [10].

Zhendong Niu and John K. Tarus conducted study to recognize the different ontology based e-learning recommended systems and prove that the use of ontology for expressing knowledge in e-learning recommender systems can bring improvement in the quality of recommendations. The methodology used is survey of the e-learning recommended system and compared and analyzed the results of various ontology based recommended systems. [11].

Kazunori Yamaguchi et.al., has developed personalized English teaching material for beginner level learners which identifies the cognitive level or difficulty level of the content in text document. The difficulty level is identified by the personalized vocabulary of the learner. The learning materials were recommended based on the vocabulary knowledge of the students. The difficulty level is determined as a ratio of the number of unknown words in a reading material. The results shows a better performance for the SVM classifiers in terms of accuracy. The research has relevance today as the number of available materials increased exponentially[12].

III. METHODOLOGY

Datasets were collected through web crawling from different e-learning websites. The contents were categorized in to three different difficulty levels namely beginner, intermediate and advanced. The webpages were parsed and stored as text files in these three folders based on its difficulty level. The dataset were created with different size to check the accuracy of the classifier at different dimensionalities. The data is preprocessed and two different feature extraction methods were applied on the data set namely Bag–of-word model and Parts of speech [POS] tagging. The resultant feature sets were made to run through

Random Forest classifier to compare the performance of the two model.

The bloom's taxonomy verbs were added to the feature set obtained after data reduction. The synonyms of each of these verbs were extracted using WordNet from NLTK took kit. The Bloom's synonyms were also added to the feature set.

A. PREPROCESSING.

Pre-processing is used to remove the noisy and unwanted data from the data set. It includes removal of Punctuations, numeric strings, tabs, stop words, white spaces, quotation marks and single letter and double letter words from the document. Documents were normalized to convert to lower case.

B. FEATURE EXTRACTION

The data obtained after preprocessing is reduced further by feature extraction methods. The feature extraction is done using Bag of Word approach and Parts of Speech Tagging.

In Bag of Word model the collection of words obtained after preprocessing is sorted to find the unique words and the frequency of each of the words.

C. PATRS OF SPEECH TAGGING

The POS tagging is used to extract the verbs from the preprocessed document. The model uses averaged, structured perceptron algorithm. It contains a pre-trained English parts-of-speech and it uses perceptron algorithms for feature extraction. It annotate a term in text with corresponding parts of speech depend on circumstances and interpretation[13]. It is normally used in Natural Language Processing for feature extraction based on verbs and nouns and other parts of speeches used in the context.

D. FEATURE SELECTION

The size of the feature set is further reduced using feature selection methods by removing less important features and taking only appropriate percentage of the total feature set. The percentage is calculated using N-Fold cross validation as follows

n = (Total No.of Feature*percentage)/100)

feature = words [0: n]

Where percentage is an integer value less than 100.

The various percentage value of the total feature set is calculated to find the best percentage of the feature selection.

The optimum accuracy is obtained when the percentage of the total data is taken between 15 to 25.The technique helps to minimize the feature size by 70 to 80 percentage of the total data.

After preprocessing the documents is divided into training and testing. 75 percentage of the data was used for training and 25 percentage was used for testing. The best training and testing percentage was obtained through N-Fold validation.

E. BLOOM TAXNOMY

Blooms Taxonomy divides a learning content into different groups based on the cognitive levels of the content[14].it divides the content as follows

Creating - Advanced level
Evaluating – Advanced level
Analyzing -Intermediate level
Applying – Intermediate level
Understanding- lowest level
Remembering –lowest level

It states that the learning at the highest level depends on the knowledge obtained in the lower levels. Therefore the concept must be remembered before understanding. To use a concept one must understand it thoroughly. Before evaluating the concept one must analyse it. To create a new concept the existing concept must be thoroughly evaluated.

Bloom's taxonomy helps to identify the cognitive level of the content with the help of different verbs used in the context.

F. MACHINE LEARNING

The supervised machine learning is used in the frame work. The training and testing data is prepared from the document obtained after feature extraction.

The training is done by using Random forest classifier. The Algorithm is chosen after comparing performance of many algorithms on the data set and the accuracy Random Forest is found to the highest in comparison with other algorithms.

G. RANDOM FOREST CLASSIFIER.

Random Forest is ensemble binary decision tree classifier. It consist of number of random decision trees. It randomly generate many binary decision trees using bagged random set of data. The trees were independent from the other and constructed using a bootstrap sample of training data [15]. The trees in the

random forest is built by adding an amount of randomness and therefore the algorithm is named as Random Forest[16] . The data left without any decision tree is named out of bag data and it is used for testing the performance of individual decision trees.
The accuracy and robustness of the RF is high in comparison with other text classification algorithms.
Its main advantages were its potential to handle over-fitting and missing data and its capability to handle large datasets without removing the variables in the feature selection and resilience to high dimensionality data, noise insensitivity, and resistance to overfitting [17] .
Random forest uses each individual tree to randomly sample from the dataset, resulting in many random trees. This is known as bagging.

The algorithm consist of many decisions trees. It uses bagging and feature randomness when building each individual trees and create an uncorrelated forest of trees whose prediction by the committee is more robust than any individual trees.

Let $d_x = \{T1, T2, T3 \dots \dots . TN \}$ be the training data set that contain N training cases. The classifier boot start the data set d_x, once bootstrapped the new training set is $e_X = \{T1, T2, T3 \dots . . TN\}$
and Ti ={ atr1,atr2,atr3 --- atrM} This means that in each training set, there exists M attributes. The classifier is constructed by randomly taking M attributes from Ti. The number of attributes in atrM is less than the total attribute. Which means atrM is equal to $\sqrt{M}$. The random forest get the most important features from the atrM to construct the decision tree.
RF chooses the best fit using the gini index

$$\text{gini(attr)} = 1 - \sum[Pj]^2 \qquad (1)$$

$$gini_{split} = \sum_{atr=1}^{n} \frac{n_{atr}}{n} gini\,(atr) \qquad (2)$$

Where Pj is the relative frequency of the feature (atr) at class J, n_{atr} is the number of randomly selected training records and atrM is the number of attributes.

Each individual tree in the random forest out put prediction and the class with more vote is taken as the prediction value.

IV. IMPLEMENTATION

The Random forest classification algorithm is made to run with datasets of different size. The algorithm is made to run separately for Bag of Words and POS tagging model. The step is repeated after adding the Blooms taxonomy verbs and synonyms to the data set.

Table 1: Random Forest Classifier in different dimensionality.

Dimensionality in Percentage	Train Time	Test Time	Accuracy
20	0.118s	0.008s	0.981
21	0.109s	0.016s	0.976
22	0.125s	0.000s	0.976
23	0.125s	0.000s	0.976
24	0.127s	0.008s	0.976
25	0.125s	0.016s	0.986

The table 1 shows the training time testing time and accuracy of the classifier in different dimensionality of data with 600 documents. The 20 to 25 percentage of the total feature is used for training the model. This percentage is taken after N-fold cross validation. The data set is giving the maximum accuracy of 0.986.

Fig 1: Classification results in different dimensionality

Figure shows the classifier result with 20 to 25 percentage of the total data set. The score, training time and testing time for different data dimensionality is shown in the graph.

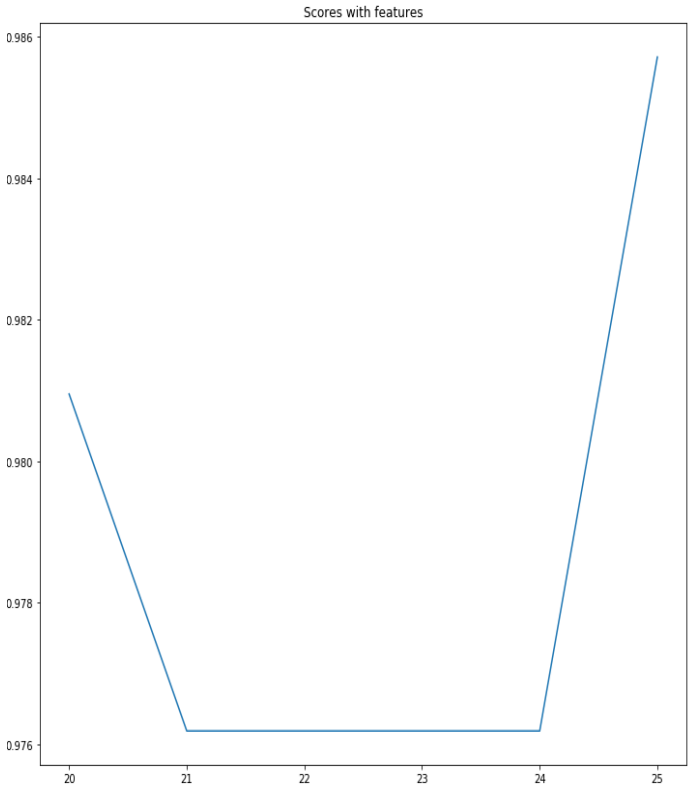

Fig 2: Classification results in different dimensionality

The line graph shows the maximum and minimum accuracy of the classifier in different data dimensionality. The maximum accuracy obtained is 0.986.

Table 2: Random Forest Classifier in different dimensionality using Bloom Taxonomy verbs.

Dimensionality in Percentage	Train Time	Test Time	Accuracy
20	0.312s	0.000s	0.992
21	0.265s	0.000s	0.983
22	0.312s	0.017s	0.992
23	0.265s	0.016s	0.992
24	0.281s	0.016s	0.992
25	0.314s	0.016s	0.992

Table shows the results of execution of the dataset on Random forest classifier with Blooms Taxonomy verbs in the feature set. The accuracy is sharply increased when the bloom verbs is used in the feature set.

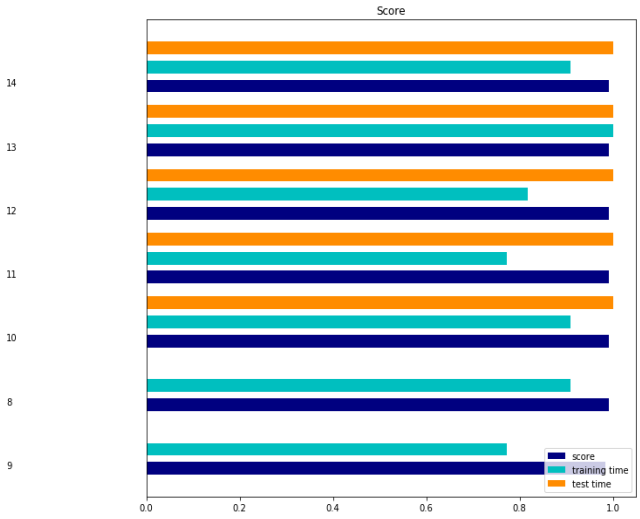

Fig 3: Classification results in different dimensionality using bloom taxonomy verbs.

Figure 3 shows the result of execution of the dataset on Random forest classifier with Blooms Taxonomy verbs in the feature set. The accuracy is sharply increased when the bloom verbs is used in the feature set.

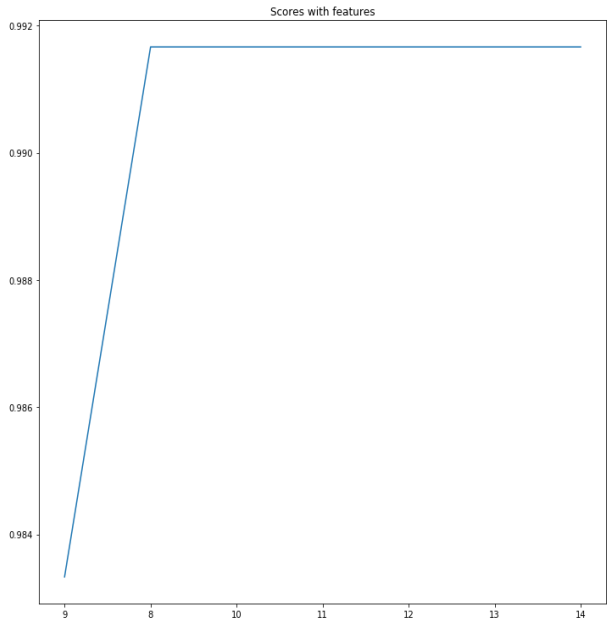

Fig 4: Classification results in different dimensionality using Blooms taxonomy verbs

The line graph shows that the accuracy of the classifier with the addition of Blooms verbs is sharply increased.

Table 3: Random Forest Classifier in different dimensionality using POS tagging.

Dimensionality in Percentage	Train Time	Test Time	Accuracy
20	0.218s	0.016s	0.975
21	0.328s	0.016s	0.983
22	0.265s	0.016s	0.983
23	0.281s	0.016s	0.983
24	0.281s	0.016s	0.983
25	0.281s	0.016s	0.983

Table 3 shows the result of Random Forest Classifier used with feature set obtained through POS tagging.20- 25 percentage of the total feature is used after N-fold cross validation. The maximum accuracy obtained is 0.983 and minimum 0.975.

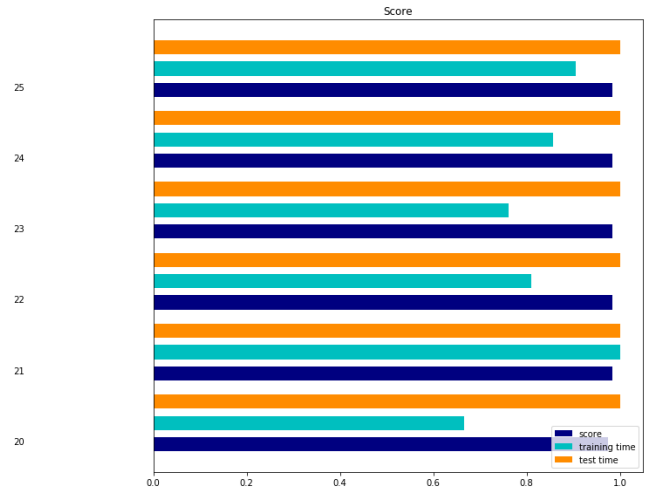

Fig 5: Classification results in different dimensionality with POS tagging.

The figure shows accuracy at different percentage of data dimensionality. The training time, testing time and accuracy were plotted as different bars in the graph.

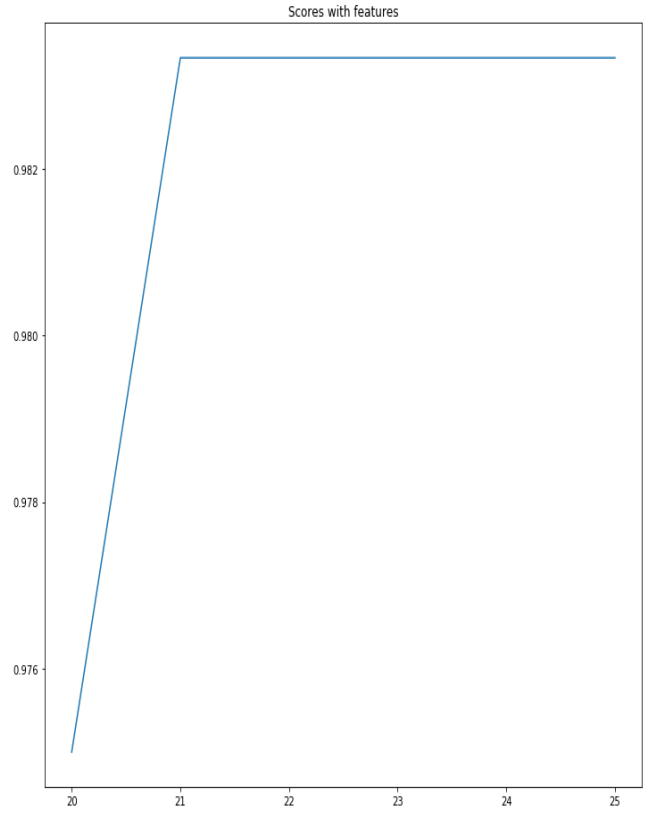

Fig 6: Classification results in different dimensionality with POS tagging.

The line graph shows accuracy at different percentage of data size. The maximum accuracy obtained is .0986

Table 4: Random Forest Classifier in different dimensionality with POS tagging using Bloom Taxonomy Verbs

Dimensionality in Percentage	Train Time	Test Time	Accuracy
20	0.343s	0.016s	0.992
21	0.281s	0.016s	0.983
22	0.281s	0.016s	0.983
23	0.343s	0.016s	0.983
24	0.328s	0.000s	0.983
25	0.312s	0.016s	0.983

Table shows the results of execution of the classifier with Blooms taxonomy verbs in dimensionality 20-25 percentage. The accuracy is sharply increased with the addition of Blooms verbs in the feature set.

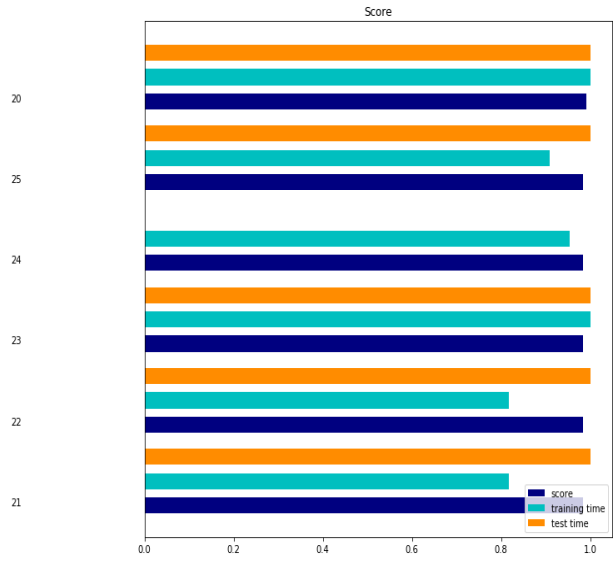

Fig 7: Classification results in different dimensionality with POS tagging using blooms taxonomy verbs

From the figure it is clear that the accuracy of the classifier with Blooms taxonomy verbs increased sharply.

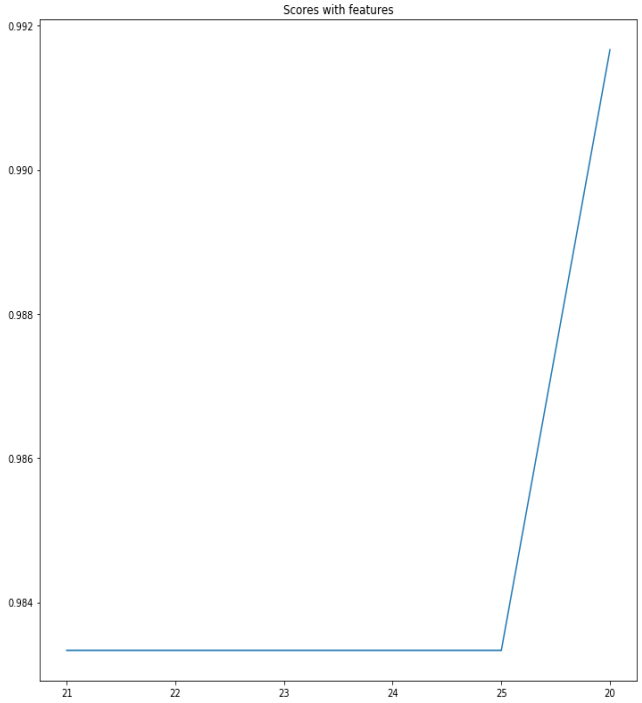

Fig 8: Classification results in different dimensionality with POS tagging using blooms taxonomy verbs

From the line graph it is clear that the accuracy of the classifier with Blooms verbs increased sharply.

V. RESULT AND DISCUSSION

The frame work is made to run with 600 files using two different dimensionality reduction methods namely bag of word model and POS tagging, using average perceptron tagger. In the POS tagger extraction method, the verbs and nouns were taken from the documents as features for training and testing the model. The results showed that the bag of word model is giving better accuracy for the classifier. Both the models were made to run using Blooms taxonomy verbs and synonyms. The accuracy of the Random Forest classifier is increased in both Bag of word model and POS tagging model when the Blooms taxonomy verbs and synonyms were added to the feature set. The reason for the higher performance with the addition of Blooms' verbs is that Random Forest works as a learning ensemble decision tree and each tree is trained on bootstrapped sample of training

data and the accuracy increases when the relevance of the added feature increases. It uses feature selection method and feature ranking based on the importance of a feature in the overall dataset[18]. Through this experimental evaluation it is found that the Blooms taxonomy verbs can be used in the classification of e-learning documents to get better performance in machine learning.

VI. CONCLUSION

The availabity of the e-learning web sites increased exponentially in recent years. The number of learners using these e-learning web sites also increased radically. There should be an intelligent method to provide right learning content to the learner to improve the learning effectives of the learner and reduces the time needed for learning. In the proposed work, Blooms taxonomy verbs and synonyms are used to improve the accuracy of the Random Forest Classification algorithm. The result shows that the Blooms Taxonomy verbs and synonyms improves the performance of the classifier, which is used to build a model to classify the e-learning contents from the web site.

REFRENCES

[1] A. A. Yahya and A. Osman, "Automatic Classification of Questions Into Bloom's Cognitive Levels Using Support Vector Machine," *Proc. Int. Arab Conf. Inf. Technol.*, no. December 2011, pp. 1–6, 2011.

[2] F. Nafa, S. Othman, and J. Khan, "Automatic concepts classification based on bloom's taxonomy using text analysis and the Naïve Bayes Classifier Method," *CSEDU 2016 - Proc. 8th Int. Conf. Comput. Support. Educ.*, vol. 1, no. Csedu, pp. 391–396, 2016.

[3] U. Fuller *et al.*, "Developing a computer science-specific learning taxonomy," *ITiCSE-WGR 2007 - Work. Gr. Reports ITiCSE Innov. Technol. Comput. Sci. Educ.*, pp. 152–170, 2007.

[4] F. Nafa, J. I. Khan, S. Othman, and A. Babour, "Mining cognitive skills levels of knowledge units in text using graph tringluarity mining," *Proc. - 2016 IEEE/WIC/ACM Int. Conf. Web Intell. Work. WIW 2016*, pp. 1–4, 2017.

[5] F. Nafa and J. Khan, "Conceptualize the domain knowledge space in the light of cognitive skills," *CSEDU 2015 - 7th Int. Conf. Comput. Support. Educ. Proc.*, vol. 1, pp. 285–295, 2015.

[6] A. A. Yahya, A. Osman, A. Taleb, and A. A. Alattab, "Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques," *Procedia - Soc. Behav. Sci.*, vol. 97, pp. 587–595, 2013.

[7] K. M. Yang, R. J. Ross, and S. B. Kim, "Constructing different learning paths through e-learning," *Int. Conf. Inf. Technol. Coding Comput. ITCC*, vol. 1, pp. 447–452, 2005.

[8] A. Nuntiyagul, K. Naruedomkul, N. Cercone, and D. Wongsawang, "Adaptable learning assistant for item bank management," *Comput. Educ.*, vol. 50, no. 1, pp. 357–370, 2008.

[9] M. K. M, S. D. H, P. G. Desai, and N. Chiplunkar, "Text Mining Approach to Classify Technical Research Documents using Naïve Bayes," vol. 4, no. 7, pp. 386–391, 2015.

[10] S. Pariserum Perumal, G. Sannasi, and K. Arputharaj, "An intelligent fuzzy rule-based e-learning recommendation system for dynamic user interests," *J. Supercomput.*, no. 0123456789, 2019.

[11] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 21–48, 2018.

[12] I. Horie, K. Yamaguchi, K. Kashiwabara, and Y. Matsuda, "Improvement of difficulty estimation of personalized teaching material generator by JACET," *ITHET 2014 - 13th Int. Conf. Inf. Technol. Based High. Educ. Train.*, 2014.

[13] R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, pp. 25–32, 2015.

[14] C. G. Johnson and U. Fuller, "Is Bloom's taxonomy appropriate for computer science?," *ACM Int. Conf. Proceeding Ser.*, vol. 276, pp. 120–123, 2006.

[15] D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, and E. Pujos-Guillot, "Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data," *Front. Mol. Biosci.*, vol. 3, no. JUL, pp. 1–15, 2016.

[16] P. Biau∗, "Analysis of a Random Forests Model G´erard," *Anaesthesiol. Intensive Ther.*,

vol. 49, no. 5, pp. 373–381, 2017.

[17] B. H. Menze *et al.*, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, pp. 1–16, 2009.

[18] W. T. Aung, K. Hay, and M. Saw, "Classification of Web Pages," pp. 372–376, 2009.