Classification of e-Textbooks based on Table of Contents Using Machine Learning Approaches

Prathibha R. J[1]., Pavan kumar S[2], Shishira G Hegde[3]

[1-3]Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, India

---***---

Abstract— Every well documented files or textbooks contain "Table of Contents" (ToC). Categorization of such documents using only the table of contents leads to elimination of processing of entire e-textbook and also reduces the time taken to identify the index terms (most significant and frequent terms in the document) that are present in the documents. Identifying the starting and ending page of ToC is a challenging task. This paper proposes a model to categorize the given document which is in Portable Document Format (pdf) using only the table of contents. The dataset containing 108 documents is used for both training and testing the proposed model. The supervised machine learning approaches like Gaussian Naïve Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbor are applied on table of contents of this dataset and obtained the accuracies of 93.87%, 83.67%, 73.45% and 82.71% respectively. The proposed system is also evaluated with the statistical measures like precision, recall and F-measure.

Keywords-Table of Content, Machine Learning Approaches, Portable Document Format.

I. INTRODUCTION

Nowadays, the web is filled with the text document in portable document format (pdf) file. All journals, textbooks, technical reports, documents etc., are available in pdf form. The main requirement for a well-defined methodology to analyze and classify pdf files has drawn many researchers' attention. The pdf files do not follow any standard rule or format, hence it makes harder for external code to manipulate and perform operations like copy and paste the content of pdf files. For this reason pdf files are rated as one of the highly unstructured formatted files. Many data mining techniques were used to classify the documents and assign each with a class or category. The existing classification approaches follow the document preprocessing, identification of index terms in the entire document and weighting of those index terms for classification. Searching for index terms in full text document consumes time and space. Hence, in this paper, we propose an effective technique to classify e-textbooks only by considering the Table of Contents (ToC) of document rather the entire document. In e-textbooks, the table of content is sufficient to identification of index terms and classification of given document. The greatest challenge here is to identify the starting and ending of table of content in pdf files that are highly unstructured and some pdf files are locked or encoded for security reasons.

This paper is organized as follows: related work on document classification techniques is presented in Section2. The proposed model explained in Section3, Section 4 gives the experimental results and discussion, and Section 5 provides the conclusion and future enhancements.

II. RELATED WORKS

Document classification is the task of assigning a document to a specific category or class. In this process representation, feature extraction and classification of documents take place. Samaneh Chagheri and Sylvie Calabretto [1] proposed a classifier to classify technical documents such as user manual and finding the structured content like table in a document. Here XML documents are considered as a tree, in which the nodes are tagged by structural labels like title, chapter, etc. The arcs represent the relation between nodes, and the leaf nodes contain the document text. Tf-idf is used for weight computation. Experiment was conducted on Reuters Corpus Volume 1(RCV1), and got 94% accuracy. Wan M.U. Noormanshah [2] presented a preliminary study based on text mining techniques: K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and decision tree on British Broadcasting Corporation dataset for classification. Another technique is proposed by Rajendra Kumar Roul [3] to classify web documents using frequent item sets that are generated by Frequent Pattern (FP) Growth algorithm which involves association analysis and vector space model is used for document representation and tf-idf is used for assigning the weights for frequent terms. Russell Power, Jay Chenet.al [4] proposed a model which mainly concentrates on feature extraction and topic specific classification.

In the recent paper by Quanzhi Li et.al [5] tried to classify the tables present in the financial document by extracting the required data from US Securities and Exchange Commission website. Their experiment is mainly based on the three important tables in the company's 10-K report. A 10-K document is an annual report required by SEC, which provides a comprehensive summary of a company's financial performance. Their target tables were Income Statement, Balance Sheet and Cash Flow. These three tables were marked in the html file by adding a special attribute to the html table element. All other types of tables were considered as Other Type. They trained the data with the features identified are row header, column header, title, etc. They trained the model with SVM classifier and got the overall accuracy of 88%.

In the paper by Mowafy Met. Al [6], the unstructured documents present in the 20-Newsgroups were used to build a phase wise classification which support both generality and the efficiency. The accuracy of 87% and 71% were achieved by Multinomial Naïve Bayes and KNN respectively. The experimental results over 20-Newgroups dataset have been validated using statistical measures like precision, recall, and F1-score.The survey paper on document classification and classifiers by Upendra Singh and Saqib Hasan [7] presents different classifiers and Natural Language Processing tasks to perform on documents. Basarkar and Ankit, presented different representation of feature vectors such as binary, tf, tf-idf to classify the given documents [8]. Authors used 814 documents of the Yahoo newsgroup dataset and got 83.1% accuracy for Naïve Bayes classifier.

While designing any Information Retrieval (IR) system, the pre-processing step is very much important. In literature, it is observed that the existing IR systems use entire content of the document for pre-processing; hence in real time this step increases the space and time complexity of the system. In the proposed work, only the Table of content of document is considered in pre-processing step for the identification of index terms, hence both time and space complexity is reduced to a greater extent which is highly essential in case of building real time applications.

III. THE PROPOSED MODEL

Architecture of the proposed model is given in Fig. 1. It contains three phases viz., i) Preprocessing phase, ii) Training phase and iii) Testing phase.The detailed descriptinon about each phase is given below.

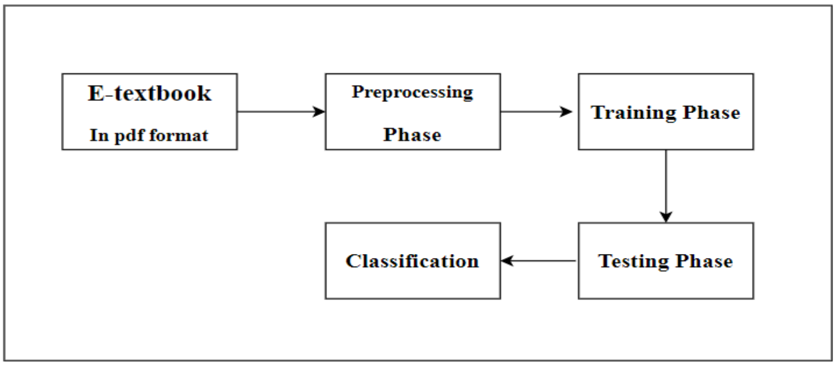

Figure 1.Architecture of the proposed model.

A. *Preproceesing Phase*

The e-textbooks belonging to different domains like physics, chemistry and some engineering courses are collected from pdfdrive.com. These e-textbooks are in PDF format.The steps involved in preprocessing phase is given in Fig. 2.

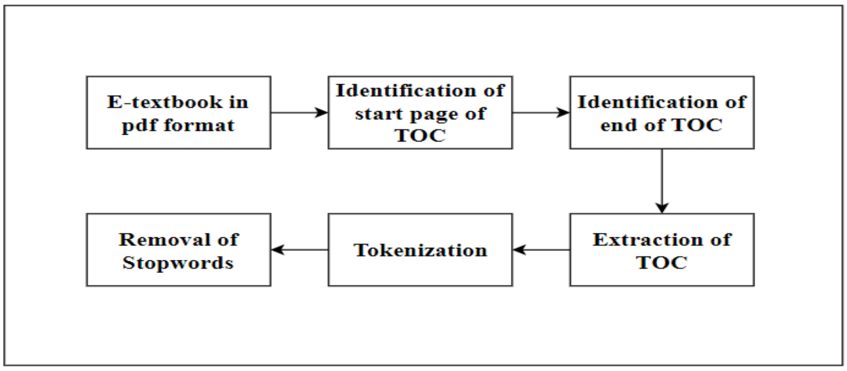

Figure 2. Steps involved in pre-processing phase.

The proposed model uses regular expressions to identify the starting page and ending page from table of content.Words that are generally used to for starting and ending of table of contents is given below. To identify these words, regular expressions used are.

- For starting page: Contents, Table of Contents.
- For ending page: Index, Bibliography, References.

After identifying the starting and ending of ToC, the entire content between starting and ending of ToC is extracted for further processing. The extracted ToC is tokenized and noises like numerals, punctuations, special characters and stopwords are removed and the remaining terms are converted into lower case for simplification.

Only the most frequently occurred words (term frequency) in the ToC are considered as feature terms. The top 30 words that are most frequently used words are considered for classification of document. On an average, the total frequent word count of ToC is around 45 for our dataset. Hand written labels are used for respective e-textbooks in the training dataset.

B. *Training Phase*

Dataset Used: The proposed work uses the dataset consisting of 108 instances with 12 classes. Each class has almost equal number of instances.

In this paper, two cases are considered in training phase viz., i) training with the content of entire e-textbook and ii) training with only the ToC of e-textbook. For comparative study, the top 30 features are trained using 4 different classifiers, SVM, KNN, Decision Tree and Gaussian Naïve Bayes. From selected dataset (160 pdf files), 64% of the dataset is used as training dataset and remaining 40% of the dataset is used as training dataset for both the cases. The steps that are used in training phase are given in Fig. 3.

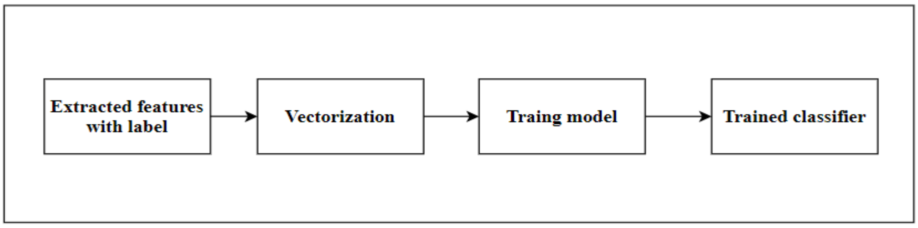

Figure 3. Steps involved in Training Phase

C. *Testing phase*

In testing phase also two cases are considered viz., i) testing with the content of entire e-textbook and ii) testing with only the ToC of e-textbook. The top 30 most frequently occurred words are extracted and considered as feature set. The features are vectorized and given to trained classifier for testing the proposed model. The classifier returns the class or the category of the input e-textbook belongs to. The steps that are used in testing phase are given in Fig. 4.

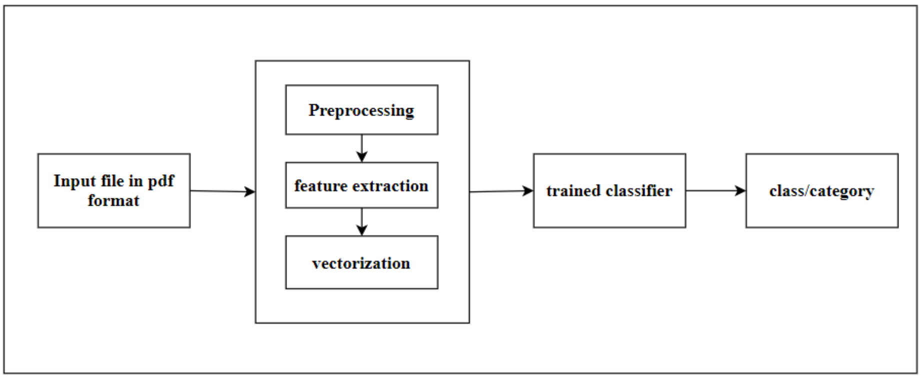

Figure 4. Steps involved in Testing Phase.

IV. EXPERIMENTAL REULTS AND DISCUSSIONS

The experimental results obtained by the proposed model for both the cases are discussed in this section. The experimental results are evaluated in terms of precision, recall and F-measure for the entire e-textbook is plotted and given in Fig. 5. From the graph, it is observed that support vector machine approach is best suitable for the classification of e-textbook by considering the content of entire document in pdf file.

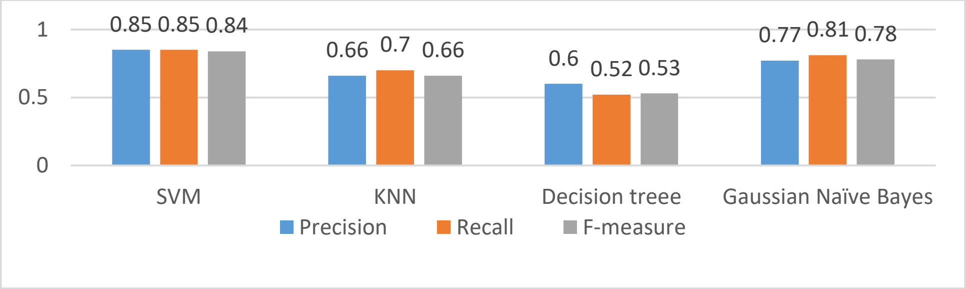

Figure 5. Performance of different classifiers on full content of e-textbooks

The experimental results are evaluated in terms of precision, recall and F-measure for only the table of content of e-textbook is plotted and given in Fig. 6. From the graph, it is observed that Gaussian naïve Based approach gave good classification of e-textbook by considering only the table of content instead of content of entire document in pdf file.

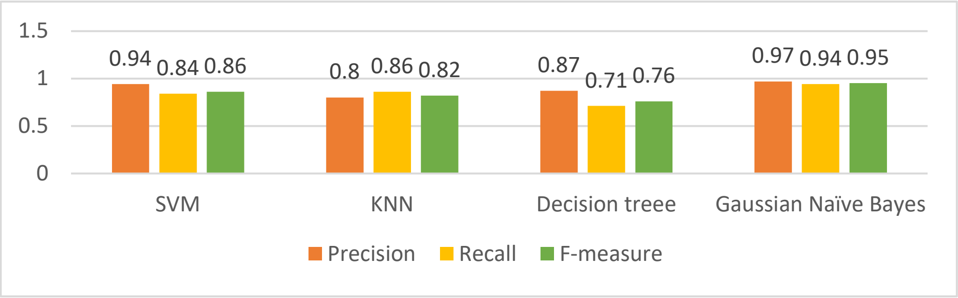

Figure 6. Performance of different classifiers using only ToC of e-textbooks

However, the time taken to classify the given e-textbook in the form of pdf file using only the table of content plays a major role in real time. The average execution time for both the cases viz.,i) classifying the document with the content of entire e-textbook and ii) classifying the document with only the ToC of e-textbook is given Fig. 6. From the Fig. 7, it is observed that the average time (in seconds) taken by the proposed classifier model using only the table of content is reduced to 1/42 times almost 97.6% and performance is also better than considering the content of entire document. It is observed from the Fig. 5 that the Gaussian Naïve Bayes classifier gave the better accuracy than SVM, KNN and Decision Tree classifiers on the selected dataset using only the table of content.

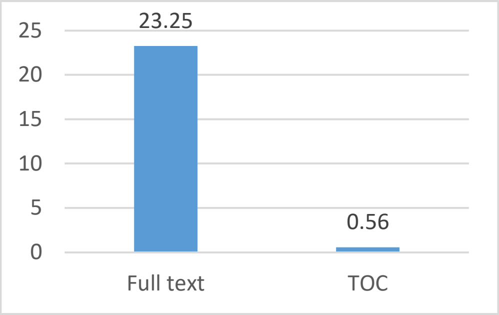

Figure 7. Average execution time taken for full text content and table of contents

Limitations of proposed model are viz., i) It works for e-textbook that are in pdf form only and ii) possibility of misclassification of given document because of the use of only table of content. In this proposed work, we have considered 108 e-textbooks in pdf form. Out of which 4, 7, 11 and 5 documents were misclassified by SVM, KNN, Decision tree and Gaussian Naïve Bayes classifiers respectively using the content of entire documents as shown in Fig. 8. There are 8, 13, 7 and 3 documents were misclassified by SVM, KNN, Decision tree and Gaussian Naïve Bayes classifiers respectively using the ToC of the e-textbook as shown in Fig. 9.

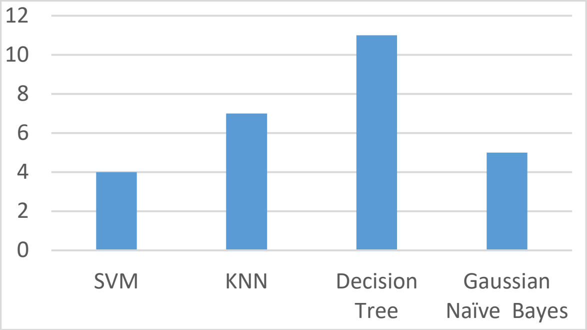

Figure 8. Number of misclassification obtained on entire content of the e-textbook.

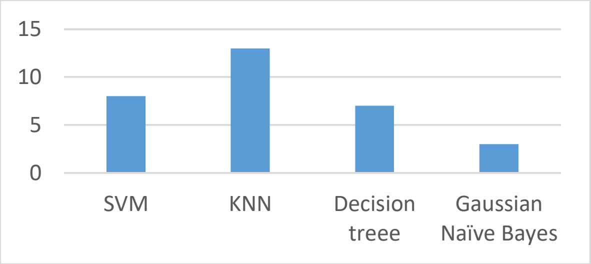

Figure 9. Number of misclassification obtained for only Table of Content

V. CONCLUSION AND FUTURE WORK

The classification of text documents using supervised machine learning algorithms by considering the content of entire document have been tried by many researchers and obtained good results. However, instead of considering the content of entire text document, if only the table of content is considered to identify the frequent index terms will give better results in text classification process. Hence, both space and time complexities can be reduced. In this paper, a text classifier is proposed which extracts only the table of content present in e-textbooks to identify frequent index terms and classifies the document using supervised machine learning algorithms. From the experimental results, it is observed that the Gaussian naïve Bayes approach gave outstanding precision of 97%, recall of 94% and F-measure of 94% on classification of e-textbook by considering only the table of content. In future, this approach can be applied on non-pdf files like doc, txt, and word files etc., for text classification.

References

[1] Chagheri, Samaneh, Sylvie Calabretto, Catherine Roussey, and Cyril Dumoulin.," Document classification: combining structure and content", 3rd International Conference on Entreprise Information Systems, 2011, pp. 8-11.

[2] Wan M.U. Noormanshah, Puteri N.E. Nohuddin, ZurainiZainol," Document Categorization Using Decision Tree: Preliminary Study" Article in International Journal of Engineering and Technology, 2018.

[3] Rajendra Kumar Roul BITS, Pilani,S.K Sahay ,BITS, Pilani, "An Effective Approach for Web Document Classification using the Concept of Association Analysis of Data Mining", 2014

[4] Russell Power, Jay Chen, Trishank Karthik, Lakshminarayanan Subramanian.: Document Classification for Focused Topics,2010 AAAI Spring Symposium Series , 2010, pp. 67-72.

[5] Quanzhi Li., Sameena Shah, Rui Fang.: Table classification using both structure and content information: A case study of financial documents, 2016 IEEE conference on Big Data, 2016, pp. 1778-1783.

[6] Mowafy M, Rezk A and El-bakry HM.: An Efficient Classification Model for Unstructured Text Document, American Journal of Computer Science and Information Technology, 2018, Vol 6, No.1:16, pp. 1-9.

[7] Upendra Singh, Saqib Hasan.: Survey Paper on Document Classification and Classifiers, International Journal of Computer Science Trends and Technology, 2015, Volume 3 Issue 2, pp. 82-87.

[8] Basarkar, Ankit. : Document Classification using Machine Learning, Thesis, San Jose University, 2017.