EDUCATIONAL DOCUMENT CLASSIFICATION USING NATURAL LANGUAGE PROCESSING

Spoorthi M, Srilekha K, Sanjana J, Kushal Kumar B N
Department of CSE, K.S Institute of Technology, Bengaluru, India
spoorthimarakkini@gmail.com, kumar.srilekha@gmail.com, sanjanaj94@gmail.com, bn.kushalkumar@gmail.com

Abstract—The paper uses supervised machine learning and content-based document classification of textual documents that are confined to four educational departments - Civil, Computer Science, Mechanical and Electrical Engineering by using TF-IDF algorithm along with Natural Language Processing for feature selection and ID3 algorithm as a classifier. The results show 80% accuracy.

Keywords---Supervised machine learning; Natural language processing; TF-IDF; Iterative Dichotomiser 3.

I. INTRODUCTION

In the present world, educational documents from various fields are increasing because of the variety of sources of information available for the students. The Manual document classification of these documents is known to be an expensive and a time-consuming job. Hence, various machine learning approaches suggest automatic construction of classifiers, that make the task easier. The main goal [1] of classification is to extract information from textual resources and to deal with operations like retrieval, classification (supervised, unsupervised and semi supervised) and summarization.

In this paper we consider supervised, content-based document classification. Content-based classification [2] is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned.

Supervised document classification is where in some external mechanism (such as human feedback) provides information on the correct classification for the documents.

Various applications of this would include topic-specific processing and information extraction, filtering and routing of documents, along with easier access to various documents that also help in management.

The numerous techniques[1] investigated in the paper are Natural Language Processing (NLP), Text classification and Machine Learning algorithms, that work together to automatically analyze, classify and discover patterns from the different types of the documents.

Natural language processing[3] is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. The various NLP operations and modules help in building better statistical machine learning algorithms. Here NLP operations are used for the data cleaning process.

Text classification (TC)[4] helps in manually building automatic TC systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules that convert expert knowledge on how to classify documents under the given set of categories. The classification task starts with a training set D = (d_1…………d_n) of documents that are already labeled with classes C_1,C_2,so on (e.g. computer, civil). The task is then to determine which class is best suitable for a new document with given data.

For example would be to automatically label each incoming document with a topic like “computers”, “computing”, or “networks” to a class labeled "Computer Science".

The remainder of the paper is organized as follows - Section 2 gives the proposed system along with the data cleaning and feature extraction, feature vector representation in the form of TF-IDF algorithm and classification in the form of the ID3 classifier algorithm. Finally we have the experimental results in Section 3. Followed by the Conclusion in Section 4.

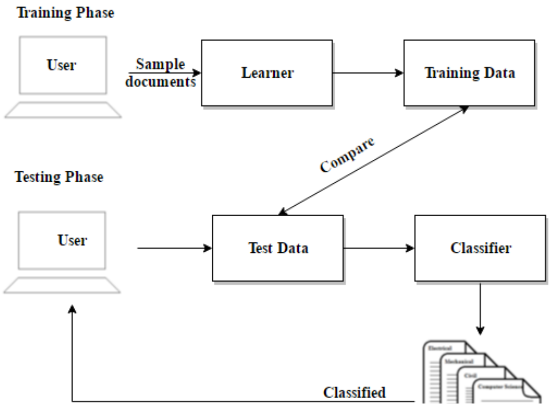
II. MATERIAL AND METHODOLOGY

A. *Proposed System*

The proposed system can mainly be divided into two distinct phases - The training phase and the testing phase.[1]

The training phase (as shown in the Figure.1) involves that the user provide a set of sample documents that belong to the four departments as mentioned in the abstract. These documents are then fed into the Learner for the creation of the training data. The Learner analyzes and preprocesses the data. The output is a model for each category which is represented by a set of classifier features that are constructed and selected with the help Natural language processing and the TF-IDF algorithm.

The testing phase involves that the user provide a new set of documents which are preprocessed to obtain a testing data set. This is then passed on to the Classifier which classifies the new documents to their specific categories by comparing the testing data set to the training data set based on the ID3 classification algorithm. The output, are a set of directories automatically created that contain the classified textual documents.



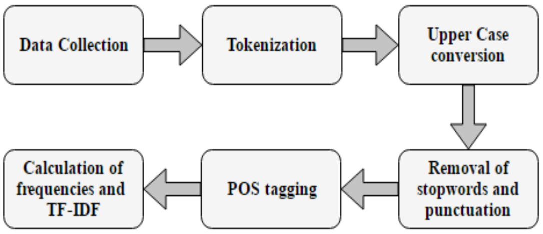

Data Collection
Tokenization
Upper Case conversion
Removal of stopwords and punctuation
POS tagging
Calculation of frequencies and TF-IDF

Figure.1.Document Classification

B. Data Cleaning and Feature Vectors

Before data mining algorithms can be used, a target data set must be assembled. A lot of unnecessary data may not actually be required for the classification process which is removed by the various processes of data cleaning as shown in Figure.2.

Figure.2.Data Cleaning and Feature extraction

For the improvement in the classification quality, we use the TF-IDF algorithm.

Tf–Idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.[1,2]**TF:** Term Frequency, measures how frequently a term occurs in a document. The normalized tf value:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).**IDF:** Inverse Document Frequency, measures the importance of a term, this is done by computing the following: IDF(t) = log_e(Total number of documents / Number of documents with term t in it).Iterative Dichotomiser 3 Classifier

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm [5] used to generate a decision tree from a dataset. The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain IG(A)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before. The algorithm uses two measures for finding the splitting attribute - Entropy and Information Gain.

a) Entropy

Entropy H(S) is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S).

$$H(S) = -\sum Pi \log_2 Pi \qquad (1)$$

b) Information Gain

Information gain IG(A) is the measure of the difference in entropy from before to after the set S is split on an attribute A.

$$IG(A,S) = H(S) - \sum P(i)H(i) \qquad (2)$$

III. EXPERIMENTAL RESULTS

AUTOMATIC EDUCATIONAL DOCUMENT CLASSIFICATION
Using natural language processing
Open

The results of the system can be summarized as follows:

A. Preprocessing

The application of NLP operations and TF-IDF algorithm show a great improvement from the naïve preprocessing techniques by producing the rare words that occur in each document for a better classification process, thereby reducing the classification time from many hours to mere minutes.

B. Classification

With around 100 documents of varying sizes, Iterative Dichotomiser 3 was able to successfully classify nearly 80%

of the documents compared to other algorithms for the given data set. According to Table.1, the number of Computer Science, Mechanical, Electronics and Civil documents correctly classified were seen to be 12, 18, 25 and 28 respectively, with around 17 documents not classified correctly.

Table.1.Document Classified

Departments	Documents Classified
C.S	12
MECH	18
EC	25
Civil	28
Not Classified	17
Total Documents	100

The Graphical representation of the document classification is shown in Figure.3 with the number of documents taken as the y axis and the various categories taken as the x-axis.

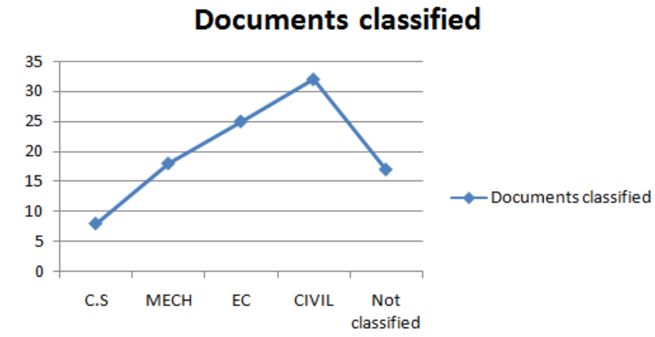


Figure.3.Number of Documents vs Categories

C. *Graphical User Interface*

The graphical user interface is highly flexible and provides the user a very easy and an interactive experience(as shown in Figure.4).

The user can either mention a directory path, that consists of the documents to be classified or select specific documents from a list of documents (as shown in Figure.5).

Figure.4.Document Classification

Once the *'classify'* button is pressed, the documents are classified and the output of the created directories are printed.

The selected documents can be viewed again, in case any document is to be added or removed.

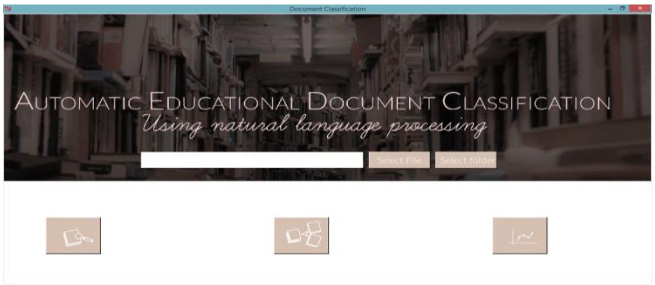

Document Classification
Automatic Educational Document Classification
Using natural language processing
Select File
Select folder

Figure.5.Page to Select Documents

IV. Conclusion

The paper deals with the classification of various educational documents. NLP operations along with TF-IDF proves to work well for preprocessing, thereby helping in the better classification of the documents using ID3. Finally the paper focuses on the development and evaluation of a system that proves to be helpful in various educational sectors of the country.

Acknowledgement

The Authors would like to thank VGST (Vision Group on Science and Technology), Government of Karnataka, India for providing infrastructure facilities through the K-FIST Level II project at KSIT, CSE R&D Department, Bengaluru.

References

[1] Bhumika1, Prof Sukhjit Singh Sehra2, Prof Anand Nayyar3, "*A Review paper on algorithms used for text classification*", Bhumika1, Guru Nanak Dev Engineering College, Ludhiana, KCL Institute of Mgmt. & Technology, Jalandhar, IJAIEM - 2013

[2] https://en.wikipedia.org/wiki/Document_classification

[3] https://en.wikipedia.org/wiki/Natural_language_processing

[4] Vandana Korde, C Namrata Mahender, Department of Computer Science&IT, Dr.B.A.M.U Aurangabad, "*Text Classification an Classifiers: A survey*", Sardar Vallabhbhai National Institute of Technology, Surat, , Department of Computer Science&IT, Dr.B.A.M.U Aurangabad, IJAIA - 2012.

[5] https://en.wikipedia.org/wiki/ID3_algorithm

Rajar aman, A.; Ullman, J. D. (2011). "Data Mining". Mining of Massiv Datasets (PDF). pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 9781139058452.

[6] Juan Ramos," Using TF-IDF to Determine Word Relevance in Document Queries", Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855

[7] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

[8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson Education,2005