Document Classification of Accreditation Documents Using Machine Learning Algorithm

Anna Fay A. Edulsa, Jocelyn B. Barbosa[1]

[1] Master Student, University of Science and Technology of Southern Philippines, Philippines

Abstract**:** *This paper presents the document classification of accreditation documents in Higher Education Institution (HEI)s specifically for Accrediting Agency of Chartered Colleges and Universities in the Philippines(AACU) accreditation. This study uses Naive Bayes algorithm in classifying documents as which area it should belongs to. For scanned documents or image-based document, the study uses Optical Character Recognition(OCR) in extracting text from the said documents. The result of the study shows 90% accuracy using cross validation.*

Keywords: *Document Classification, Naive Bayes, Optical Character Recognition*

1. Introduction

Accreditation requires evidences if the standards are being met of a particular program in the HEI that is being visited. These evidences are in the form of documents, videos and/or ocular visits. In each HEI, there are accreditation tasks force that assigned to gather, prepare and submit documents. The assigned accreditation task force then manually classify a document as to which area it belongs.

These documents are in the forms of Portable Document Format(PDF), Document File Format(Doc) and Scanned PDF. Scanned PDF is a typical example, sometimes it looks like the normal PDF file created from Word, but actually when you scan a paper using a scanner, the whole content will be captured as an image [Lighten Software Inc.]. If you want to convert the scanned pdf into text, you have to use Optical Character Recognition (OCR) technology.

OCR is a widespread technology to recognize text inside images, such as scanned documents and photos. OCR is the mechanical or electronic conversion of images typed, handwritten or printed text into machine-encoded text (A. Chaudhuri et al, 2017).

Document Classification is one among the emerging research area in Text Mining. It is well proven approach to organize the huge volume of textual data. The solutions to the most of the text mining applications are solved by using machine learning algorithms (Arivoli et. al,2015).

In the field of text mining, a number of approaches have been developed to represent and classify text documents (C.Saranya Jothi et, al.,2015).

According to Ankit Basarkar in their study on Document Classification Using Machine Learning, to perform document classification algorithmically, documents needs to be represented such that it is understandable to the machine learning classifier. The report discusses the different types of feature vectors through which document can be represented and later classified. Their conclusion, is that Term-Frequency, Inverse-Document Frequency (TfIdf) should be the preferred vectorizer for document representation and classification.

Nowadays, there are existing studies on document classification. An Efficient Classification Model for Unstructured Text Document supports the generality through following the logical sequence of the process of classifying unstructured text documents step by step. The experimental results over 20-Newgroups dataset have been validated using statistical measures of precision, recall, and f-score (Mowafy M et. al, 2018).

This study elaborates the use of TfIdf with Multinomial Naive Bayes algorithm. However, unlike the proposed study it compares the use of TfIdf with Multinomial Naive Bayes and TfIdf with K-nearest Neighbor. This study concludes the superiority of using Multinomial Naive Bayes with TfIdf than K-nearest Neighbor as an approach for text document classification.

Automatic Educational Document Classification Using Natural Language Processing use a 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. This study involves the usage of TfIdf algorithm along with natural language processing for feature construction and selection (Spoorthi M et. al, 2017).

In contrast with the proposed study, it classifies accreditation documents rather than four educational departments.

Semantic Analysis of Accreditation Document Using Ontology focuses on the development of the domain ontology which includes document extraction and classifying document to its area and even sub-areas of the exhibits, developing an algorithm and evaluating its efficiency and evaluating the prototype by the respondents using ISO 9126.

The researcher developed the ontology by writing down all the related terms based on each area using the old exhibits. When a new document is feed in the system, keyword density algorithm is used to determine which keyword has the most density in a document as a base to determine the main idea of such document (Alejandria et. Al, 2017).

Differently, the proposed study uses TfIdf feature extraction method to get the keywords associated to a certain area. Once extracted, a machine learning algorithm, naive bayes will be used in order to classify the documents.

Utilizing Image-Based Features in Biomedical Document Classification uses both image- and text-based features in order to identify articles of interest, in this case, pertaining to cis-regulatory modules in the context of gene-networks. Our approach is novel and different use OCR to extract individual characters (Kai Ma et. al, 2015).

Another study is Exploring a New Space of Features for Document Classification: Figure Clustering explored image clustering as basis for constructing visual words for representing documents. Once such visual words are formed, the standard bag-of-words are formed along with commonly used classifier, such as the Naive Bayes, can be used to classify a document(Chen et. al.).

Unlike the proposed study, the proponent focuses on classification tasks of identifying documents pertaining to academic accreditation rather than biomedical documents.

Naive Bayes Classifier is the simple Statistical Bayesian Classifier (Archana and Elangovan, February 2014). It is called Naïve as it assumes that all variables contribute towards classification and are mutually correlated. This assumption is called class conditional independence (BhaveshPatankar and Chavd, December 2014). Its advantages are it requires short computational time for training, improves the classification performance by removing the irrelevant features and has good performance (Jadhav and Channe, 2014).

The proposed study is a desktop application what will allow the accreditation assigned tasks force to upload both text-based and image-based documents to the system as most of the existing accreditation documents are in these file formats. The proponent will use Optical Character Recognition (OCR) technology in converting the image-based documents in to text.

The system will then classify as to which area those documents should belong to. The proponent will use Naive Bayes classification method. This study will be implemented in the Department of Information Technology in the University of Science and Technology of Southern Philippines(USTP), Cagayan de Oro City, Misamis Oriental, 9000 and for AACCUP accrediting body only.

2. Review Of Related Literature

2.1. An Efficient Classification Model for Unstructured Text Document

The aim of this paper is to present a classification model that supports both the generality and the efficiency. It supports the generality through following the logical sequence of the process of classifying the unstructured text documents step by step; and supports the efficiency through proposing a compatible combination of the embedded techniques for achieving better performance. The experimental results over 20-Newgroups dataset have been validated using statistical measures of precision, recall, and f-score. The results have proven the capability of the proposed model to significantly improve the performance (Mowafy M et. Al, 2018).

2.2. Automatic Educational Document Classification Using Natural Language Processing

In this paper we use a 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. This study involves the usage of TF-IDF algorithm along with natural language processing for feature construction and selection. It also investigates ID3 (Iterative Dichotomiser 3) algorithm as a classifier for the given data set. The results give us 80% accuracy (Spoorthi M et. al, 2017).

2.3. Semantic Analysis of Accreditation Document Using Ontology

The researcher used evolutionary prototyping wherein the developers learned from the user a more accurate end products that allowed flexible design and development. The prototype was developed using different tools for extraction, especially Optical Character Recognition (OCR), ASP.NET as front-end and SQL Server for the database, as well as, structuring its own ontology. Unstructured files are basically the input for the system. A dashboard is also developed to monitor the number of documents in all areas of accreditation (Alejandria and Vinluan, May 2017).

2.4. Utilizing Image-Based Features in Biomedical Document Classification

Extending on our new idea, which we have recently introduced, of using OCR-based features to identify DNA contents in images, we combine image and text based classifiers to categorize documents as relevant or irrelevant to cis-regulatory modules. Using a set of hundreds of articles, marked by experts as relevant or irrelevant to cisregulatory modules, we train/test image and text based classifiers, as well as classifiers integrating both. Our results indicate that the latter show the best performance with Recall, F-measure and Utility measures all above 0.9, demonstrating the significance of incorporating image data, and specifically OCR-based features, into the document categorization process (Kai Ma et. al, 2015).

3. Methodology

This chapter discuss the methodology of the study.

3.1. Framework

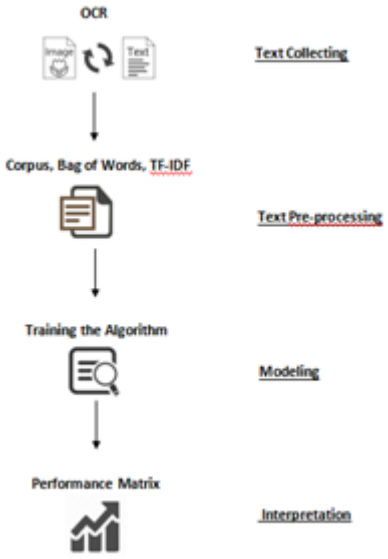

Fig. 1. Framework of Document Classification of Accreditation Documents Using Machine Learning.

3.1 Text Collecting

This phase is where we build our corpora or we prepare our data set. We collect text from both text-based and image-based documents.

3.2 Text Preprocessing

Bag-of-words uses count of word occurrence as a feature. However, not all words which appear more often should have a greater weight in textual analysis. Words such as "the", "will" and "you" are of very little significance. Instead, the words which are rare are the ones that actually help in distinguishing between the data, and carry more weight.

In this study, we will use TF-IDF technique. Term frequency-inverse document frequency (tf-idf) can be used to down weight those frequently occurring words in the feature vectors.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Fig. 2. TF-IDF Formula

Below is a snippet of document-term matrix containing top 30 for Area I and Area II containing TF-IDF values.

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Fig. 3. TF-IDF values

3.3 Train n/Test Split and Cross Validation

Once we build our data set, we then split it into training data and test data.

3.4 Modeling

Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction (Rascha, 2014).

	0	1	2	3	4	5	6	7	8
	appointment	city	college	contract	development	education	employee	entitled	faculty
Doc1	0	0	0	0	0.32577747	0.58495	0	0	0
Doc2	0.65431048	0.2043111	0	0	0	0	0	0	0
Doc3	0	0.1292535	0.1370713	0	0	0.13707	0	0	0
Doc4	0.41486659	0.1295438	0.2883056	0	0	0	0	0	0.24743387
Doc5	0	0	0	0	0.49685011	0.5269	0	0	0
Doc6	0	0	0	0	0.46382591	0	0	0	0
Doc7	0	0	0.225043	0	0.21220766	0.38103	0	0	0.23939163
Doc8	0	0	0	0	0.30049985	0.31868	0	0	0.57396694
Doc9	0	0.1208546	0.1281645	0	0.1208546	0.217	0	0	0
Doc10	0	0	0	0	0	0	0	0	0
Doc11	0	0	0	0	0.44177733	0	0	0	0
Doc12	0	0	0	0	0	0	0	0	0
Doc13	0	0.1221091	0.2192538	0	0	0	0	0	0
Doc14	0	0.5082202	0.4348294	0	0	0	0.369555	0	0
Doc15	0	0	0	0.58041	0	0	0.474645	0.44572889	0
Doc16	0.41486659	0.1295438	0.2883056	0	0	0	0	0	0.24743387
Doc17	0	0	0	0	0	0	0.143499	0.37948864	0
Doc18	0	0	0	0	0	0.69453	0	0	0.52420735
Doc19	0	0.2099415	0	0	0.2602171	0	0	0	0.33379171
Doc20	0	0.1653947	0.1753986	0.39422	0.20500248	0.30518	0	0	0.28755556

Fig. 4. Naive Bayes Formula

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \ldots P(x_n|c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x|c)$$

4. Results And Discussion

The study uses cross validation in testing the accuracy of the model. It results 90% accuracy.

According to Ashari, et.al (2013) on their study on Performance Comparison between Naive Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, based on Percision, Recall, Fmeasure, Accuracy, and AUC, the performance of Naïve Bayes is the best.

5. Conclusion

Optical Character Recognition help the researches extract terms from scanned documents. Term Frequency Document Inverse Frequency plays a very important role in preprocessing text documents. Naive Bayes algorithm has a very important role in classifying the documents to their respective areas.

References

[1] Lighten Software Inc. How can you distinguish scanned PDF from normal PDF file? Available from: https://www.lightenpdf.com/knowledge-base/scanned-pdf-ocr.html (12.11.2018)

[2] Mowafy M. et. al. (2018) An Efficient Classification Model for Unstructured Text Document, American Journal of Computer Science and Information Technology

[3] Ankit Basarkar (2017) Document Classification Using Machine Learning, San Jose State University Scholar Works

[4] A. Chaudhuri et. al.(2017) Optical Character Recognition Systems for Different Languages with Soft Computing, Studies in Fuzziness and Soft Computing 352, DOI 10.1007/978-3-319-50252-6_2

https://doi.org/10.1007/978-3-319-50252-6_2

[5] Mark Cherwin et. al.(2017) Semantic Analysis of Accreditation Document Using Ontology, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 5

[6] Spoorthi M. et al. (2016) Automatic Educational Document Classification Using Natural Language Processing International Journal of Engineering Trends and Technology Volume 35, No. 4

[7] P.V Arivoli et. al.(2015) Document Classification Using Machine Learning Algorithms - A Review. Internation Journal of Scientific Engineering and Research ISSN : 2347-3878

[8] C. Saranya Jothi, et. al.(2015) Machine Learning Approach to Document Classification using Concept based Features, International Journal of Computer Applications, Volume 118, No.20

[9] Kranti Ghag et. al(2014) SentTFIDF - Sentiment Classification Using Relative Term Frequency Inverse Document Frequency, International Journal of Advanced Computer Science and Applications, Vol. 5, No.2

[10] Ahmad Ashari et. al. (2013) Performance Comparison between Naive Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, International Journal of Advanced Computer Science and Applications Volume 4, No. 11.

[11] Ravina Mithe et. al.(2013) Optical Character Recognition, International Journal of Recent Technology and Engineering ISSN: 2277-3878, Volume 2, Issue 1.