Automated Categorization of Research Papers with MONO Supervised Term Weighting in RECApp

Ivic Jan A. Biol[1], Rhey Marc A. Depositario[2], Glenn Geo T. Noangay[3], Julian Michael F. Melchor[4], Cristopher C. Abalorio[5], James Cloyd M. Bustillo[6]

Computer Science Program, Computer Education Department,
ACLC College of Butuan, Butuan City, Philippines[1, 2, 3, 4, 5, 6]
Caraga State University, Butuan City, Philippines[5]
Graduate Programs, Technological Institute of the Philippines, Quezon City[5, 6]

***Abstract*—Natural Language Processing, specifically text classification or text categorization, has become a trend in computer science. Commonly, text classification is used to categorize large amounts of data to allocate less time to retrieve information. Students, as well as research advisers and panelists, take extra effort and time in classifying research documents. To solve this problem, the researchers used state-of-the-art supervised term weighting schemes, namely: TF-MONO and SQRTF-MONO and its application to machine learning algorithms: K-Nearest Neighbor, Linear Support Vector, Naive Bayes Classifiers, creating a total of six classifier models to ascertain which of them performs optimally in classifying research documents while utilizing Optical Character Recognition for text extraction. The results showed that among all classification models trained, SQRTF-MONO and Linear SVC outperformed all other models with an F1 score of 0.94 both in the abstract and the background of the study datasets. In conclusion, the developed classification model and application prototype can be a tool to help researchers, advisers, and panelists to lessen the time spent in classifying research documents.**

Keywords*—Text classification; supervised tern weighting schemes; optical character recognition; machine learning algorithms

I. Introduction

Writing and publishing have taken popularity on the internet using online services where text classification plays an important role [1]. An example where text classification can be applied is in the increasing amount of published research documents online or offline due to the advancement of computer and information technologies [2]. Documents, in this case, refer to textual records, and each copy contains a group of words that ranges from sentence to paragraph long. It is through the use of text classification, the prediction and classification of documents can be made possible by categorizing them into which class they belong based on their inherent properties [3].

While there are many ways to classify research papers online, there is also a need to categorize those with only physical copies. Approved research papers refer to peer-reviewed and panel-evaluated complete research in local school libraries. Additionally, research papers still in the proposal period are subject to revisions and need to be more easily distinguishable whether they are suitable for the course. Moreover, classifying large documents is time and energy-consuming [4]. For such reasons, it is necessary to make a tool that efficiently organizes approved and work-in-progress research papers individually or in bulk.

In this study, to classify research documents, first, OCR will be used. Optical Character Recognition (OCR) acquires an image through the use of a device, usually a camera or scanner, and then converts it to digital text [5][6][7]. Then, supervised term weighting schemes are applied to assign a weight for each term in every document, enhancing text classification performance [8]. The documents will be assigned to their designated classes using different machine learning algorithms [9].

The researchers came up with the idea to design a classification model using different combinations of supervised Term Weighting Schemes (TWS) and Machine Learning Algorithms, as well as prove which combination of Supervised TWS and Machine Learning algorithms is the fastest and yields the results with the highest accuracy for classifying our dataset. Finally, this study will develop an application that will use the designed classifier model to categorize research papers while allowing users to import text images or use the device's camera for text image acquisition.

The subsequent sections of this paper cover various important aspects of the study, including: a literature review on text classification, OCR, term weighting schemes, and machine learning algorithms in Section II, an in-depth examination of the research methods used in the study in Section III, a presentation of the results and discussion of the findings in Section IV, and a conclusion summarizing the key takeaways and recommendations for future research in Section V.

II. Literature Review

In this section, we present a comprehensive review of relevant literature pertaining to this study. This includes a detailed examination of research on Text Classification, Optical Character Recognition (OCR), Weighting Schemes, and various machine learning algorithms. The sub-sections will thoroughly understand the field's current state and serve as a foundation for this research.

A. Text Classification

Text classification or also known as text categorization, is a technology for information organization and management wherein it has been proven to be effective and efficient [10]. In natural language processing, text classification is a crucial task and has been its foundation. Over the years, numerous research in this area has been published due to the unprecedented success of deep learning [11].

Through the use of the term frequency-inverse document frequency (TF-IDF), Latent Dirichlet Allocation (LDA), and K-Means clustering, [2] designed a research paper classification model wherein it classifies and clusters similar papers based on its abstract. First, a keyword dictionary which contains groups of keywords that have similarity in meaning, is constructed into one representative keyword or topic. However, yielding results will still cause high running time. In order to solve this problem, they used LDA to extract topic sets before calculating the word and document frequency using TF-IDF. Then, the set of data is classified into classes based on their similarity using K-means clustering, a clustering technique used to minimize distances between every data point and the nearest cluster or centroid. Finally, to evaluate the accuracy of their classification system, they used F-score, an evaluation metric used in text classification, combining precision and recall values.

B. Optical Character Recognition

Over the past few decades, the area of pattern recognition has been a topic of study which is known as Optical Character Recognition. For many diverse styles of programs in different fields, Optical character recognition is the bottom of it, which we use in our daily life. These days, Optical Character Recognition is being used in many different areas of research [12][13].

P. Divya et al. (2021) developed a web-based optical character recognition application using flask and tesseract [14]. Their website allows users to upload image format data and convert it into machine-editable text in a fraction of a second. It can also read and convert handwritten data with a slightly lower accuracy compared to digitized text. Their OCR model has been proven to be accurate by testing a combined handwritten and digitized test set of 1000 images. In their study, they found that the best input that their system can convert to text is digitized black and white with an accuracy rate of 98%.

C. Supervised Term Weighting Scheme

The selection of an appropriate term weighting scheme is important in text classification tasks as it has significant effects on its performance. Term Weighting Schemes (TWS) determine how texts would be represented in the vector space model. The terminology Supervised Term Weighting Schemes have been gaining popularity in the recent years while term frequency-inverse document frequency (TF-IDF) is still widely used. But its disadvantage is that it does not train text using the available categories [15][16] and is considered an unsupervised term weighting scheme, unlike the state-of-the-art supervised term weighting schemes: TF-IGM and TF-MONO.

Most of the popular novel supervised TWS focuses on assigning weights based on how the terms occur throughout the classes, which is a proven and effective method in term weighting. However, Dogan and Uysal (2020) believed that this information may not be enough to determine the terms' power in the document [17]. They proposed the novel STW scheme Term Frequency Max-Occurrence and Non-occurrence (TF MONO), which makes use of the non-occurrence information along with the max-occurrence information of terms in the document.

TF-MONO is a supervised TWS that uses the class with the max-occurrence as well as the non-occurrence information in the document frequency. The procedure of the MONO TWS is represented in Fig. 2. Fig. 1 illustrates the MONO TWS designed by [17] separated into seven (7) steps, and visualized by [18][19], which will be further explained in details.

The steps for performing MONO on a text collection are as follows:

1) Sort the document frequency of a term in descending order.

2) Divide the sorted document frequency into two groups: one for the highest class document frequency values and the other for the rest of the classes.

3) Represent the first group with a max-occurrence (MO) ratio and the second group with a non-occurrence (NO) ratio.

4) Calculate the MO ratio as the ratio between the quantity of text documents in the class where the term occurs most and its total quantity of text documents.

5) Calculate the NO ratio as the ratio between the quantity of text documents in the rest of the classes where the term does not occur and the total quantity of text documents in the rest of the classes.

6) Calculate the product of MO and NO ratios and assign it as the MONO (Local) weight of the term.

7) Calculate the MONO (Global) weight of the term by using the MONO (Local) weight and a balance parameter α with a default value of 7.0.

8) Finally, two (2) term weighting schemes based on MONO (Global) collection frequency factor are shown.

D. Machine Learning Algorithms

As the internet expands, the number of unorganized data is also increasing. Thus, intelligent programs that use machine learning in classifying documents have been researched and developed to efficiently access information. Some of the machine learning techniques used for document classification include Naïve Bayes, Support Vector Machine, Decision Trees, etc. [20]. A. Barua et al. (2021) used the most common machine learning methods, namely: Logistic Regression, Support Vector Classifier, Decision Tree(C4.5) [24], Naïve Bayes, Random Forest, and K-Nearest Neighbor, to classify articles about sports into four (4) different categories: Cricket, Football, Tennis, and Athletics. 80% of the data were used for training, while the other 20% were for testing [21]. Using F1-score for evaluation, the result shows that the "Cricket" category has the highest F1 value as it has the most data used

for training among all the classes. Meanwhile, the category with the lowest F1 value is "Athletics" due to its low number of training data. Regarding the machine learning models, Naïve Bayes has the best performance (98.53%) for identifying documents in the cricket class with unigram + bigram + trigram feature while KNN has shown poor performance on an imbalance dataset.

III. METHODOLOGY

This section outlines the key components integral to the research study, which employs machine learning models for data analysis. These include the data sources, research design, data collection and preprocessing techniques, machine learning models, evaluation metrics, and the application simulation used to test the models. All these components work together to provide a comprehensive framework for conducting a thorough and systematic study, utilizing machine learning models to extract insights from data, and evaluate the performance of the models. The research framework and its implementation are illustrated in Fig. 1, providing a clear understanding of the methodology adopted in the study.

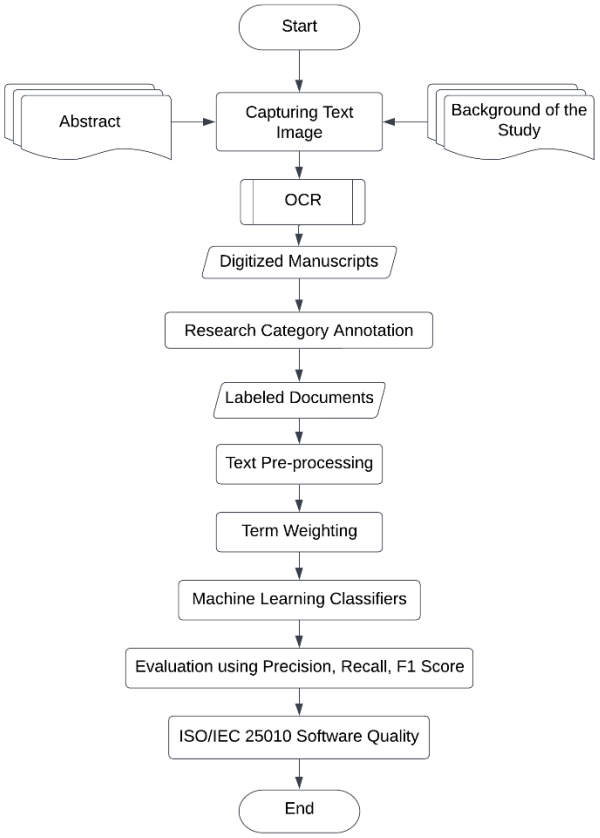


Fig. 1. Framework of the study

The *components* in the framework of the study includes the following:

A. Data

The data used in this study is gathered from the Abstract and Background of the Study (BOS) of research documents from ACLC College of Butuan, Saint Joseph Institute of Technology, and Caraga State University. A total of 462 research documents were gathered from the mentioned local schools. To address the lack of data, the researchers used online websites as a source of additional data, with a total of 519 research documents. However, the dataset is still imbalanced, with the feasibility study having only 99 documents. In order to make the dataset unbiased, oversampling is performed on the feasibility study, adding 99 more documents. Overall, there are a total of 1,121 research documents with four (4) categories: Capstone (313), Thesis (306), Case Study (304), and Feasibility Study (198).

B. Converting Research Papers to Image Format

Research papers acquired online are in .pdf image format. However, numerous research papers from local schools are not digitized. Therefore, it is necessary to capture images of the research documents' title, abstract, and background of the study, the parts of a research document necessary to be used as our data. The image format of the said images is .jpeg and can still be converted to digital text through OCR.

C. PyTesseract OCR

The images were processed one by one through a python script that utilizes PyTesseract, a well-known python Optical Character Recognition library, to convert text images into a digital format. This library allows for accurate and efficient conversion of images to text, making it an ideal choice for this task. The script iterates through all the images, performs the conversion and saves the digital text in a file, ready for further analysis.

D. Labeling Research Categories

The converted texts were then placed in a csv file for easy access and organization. The Pandas python library was used to facilitate this process, as it allows for efficient manipulation and storage of large datasets. Each text was labeled manually by researchers according to its predefined category to aid in the analysis and classification process.

E. Pre-Processing

Text pre-processing is an integral part of text classification as it can improve the overall quality of a dataset. It can clean the dataset by removing unneeded parts of the text, such as repetitions and spelling errors [22]. Text pre-processing includes four (4) basic processes: tokenization, stop words removal, stemming, and vector space model [23]. Tokenization involves the removal of spaces and taking unique words from the document. Stop word removal is the process of removing stop words which are prevalent words with little to no meaning in the document. Stemming converts all of the words in the document into their root words to reduce the unique words in the document. Vector space modeling, also known as vectorization or term weighting, assigns weight to unique words. To prepare the dataset for this study, the researchers preprocessed the text in the following order:

1) Removing all punctuations and transforming all characters to lowercase to standardize the text and make it easier to work with.

2) Tokenization was applied to split the text into multiple words which makes it more manageable for analysis.

3) Stop words were removed as they do not carry any significant meaning for the analysis.

4) Filtering the features by length was done to remove random words that hold no meaning, this helps to reduce the noise in the dataset.

5) Stemming was applied using Porterstemmer which reduces words to their base form, this helps to group similar words together, and enables better analysis.

These steps helped to clean the dataset, making it more organized and ready for the analysis. The preprocessing steps not only standardize the text but also make it more manageable and focused for analysis, which ultimately results in more accurate and meaningful results.

F. Term Frequency Distribution

In order to apply any term weighting scheme, it is crucial to first calculate the term frequency (TF) of each term in the dataset. This provides an understanding of the significance of each word across all documents in the dataset. However, simply counting the term frequency does not take into account other important factors, such as the occurrence of specific terms in certain categories. As a result, it is necessary to consider additional information to accurately calculate the weight of each term.

G. Supervised Term Weighting Scheme

Supervised Term Weighting Schemes such as TF-MONO and its square root variant, SQRT TF-MONO was used to assign weights to each term in the dataset. This helps to improve the performance of text classification by transforming the text in documents into vectors in the vector space (Feng et al., 2018). A balance parameter α with a value of 6.0 was introduced to compute the global term weight value for each term.

H. Machine Learning Algorithms

The researchers utilized state-of-the-art machine learning algorithms to categorize the dataset. The data was split into two (2) sets: 70% for training and 30% for testing. The researchers then applied three (3) machine learning algorithms, namely Naive Bayes, K-Nearest Neighbors, and Linear Support Vector Classifier, in combination with supervised term weighting schemes to compare their performance. The aim was to identify which combination of algorithm and term weighting scheme yields the best results.

I. Evaluating the Classification Model

The classification model's performance will be evaluated by calculating precision, recall, accuracy, and F1 score.

Precision - measures how many positive predictions are correctly predicted.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Recall - measures how many true positive cases the classifier correctly predicted over the total number of positive case.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

F1 Score – measures by combining both recall and precision. Also known as the harmonic mean of the two, calculating their average.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

J. Application Simulation

The researchers chose to use Python as the primary programming language for the back-end of the study due to its interactive, object-oriented, and high-level nature. Additionally, Python offers a wide range of available modules and libraries, including the Scikit library, which makes it simple, flexible and dynamic in functionality. This made it an ideal choice for the implementation of the study, including the development of the RECApp application. Furthermore, Python is a popular choice among researchers and developers because it is easy to learn, has a large community support, and is widely used in various applications such as web development, machine learning and data analysis.

K. ISO/IEC 25010 Software Quality

This software application will be evaluated in accordance with the ISO/IEC 25010 Software Quality standards, which provide internationally recognized guidelines for assessing the quality and performance of software systems.

IV. RESULTS

The researchers conducted an evaluation of the effectiveness of PyTesseract OCR by selecting sample images of abstracts and backgrounds of the study from various research documents in their dataset. They performed OCR on the images and pre-processed the data to clean it.

Fig. 2 illustrates the visual representation of the abstract of a thesis titled "Safewatch: A Quarantine Symptom-Monitoring Web Application with Knowledge Discovery Using Apriori Algorithm and Naive Bayes Classifier" and the results obtained after performing OCR on it.

Fig. 3 illustrates the visual representation of the background of the study for a capstone project titled "Barangay Information System with Decision Support System of Barangay Baan km.3 Butuan City" and the results obtained after performing OCR on it.

Fig. 4 shows the results of the pre-processing techniques employed in the extracted text of abstract and BOS. Table I illustrates the five terms with the highest frequency in the abstract dataset, along with their respective occurrences in each category.

Fig. 5 presents a word cloud comprising 200 of the most frequently occurring terms in both the Abstract and BOS datasets, arranged from left to right.

Table II presents the global factor term weights of the most frequently occurring terms in both the Abstract and BOS datasets. These values are generated from the computed results of the MONO Global method, taking into account an alpha value of 6.0. These global values will be multiplied with the term frequency if TF-MONO is to be calculated and the square root of the term frequency if SQRTF-MONO is to be calculated. It can be seen that the feature "system" has the

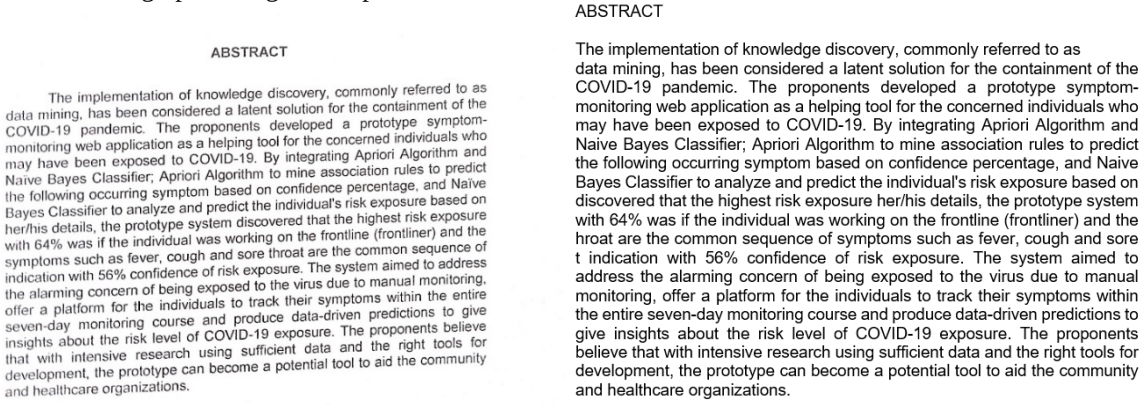
highest global weight of 0.029, indicating its prevalence in the abstract dataset. In contrast, the feature "custom" - which likely refers to stemmed customers - occurs more frequently in the BOS, suggesting that it is more prevalent in this dataset.

Table III displays the precision and recall scores for various combinations of term weighting schemes and machine learning algorithms used in this study for the Abstract dataset. The results indicate that the model using the squared root of TF-MONO trained with LinearSVC consistently performed the best, with a precision of 0.94, recall of 0.94, and F1 score of 0.94. However, as shown in Table 4, the SQRT TF-MONO model trained with MultinomialNB achieved slightly higher precision and recall scores, with values of 0.91 and 0.86, respectively. Additionally, the F1 score of the LinearSVC model still outperformed the classification performance of all five other trained models.

The researchers evaluated the application by conducting a survey that adhered to the ISO/IEC 25010 Standards. Twenty-five (25) participants completed the survey, comprising 15 questions divided into sections, each containing a minimum of 3 and a maximum of 4 questions. The results, displayed in Table IV, show the average percentage of responses for each section and are rated on a scale of 5 to 1, with 5 indicating strong agreement and 1 indicating strong disagreement.

V. DISCUSSIONS

The researchers developed an application that can classify the category of imported research papers based on four predefined categories. The application uses OCR to read text on image or pdf files and supports the classification of single or multiple research papers, and minor inaccuracies are seen in Fig. 2 and 3. Pre-processing was done in five steps to alleviate this problem, and the results are shown in Fig. 4, along with the topics extracted from the research document. The number of occurrences of each word was shown in Table I to prove the relevance of the words to each category. The results generated in Tables III and IV show that all combinations of supervised term weighting schemes and machine learning algorithms have high F1 Scores. According to the survey results, the majority voted "strongly agree" on all characteristics of ISO/IEC 25010 Standards Characteristics (see Table V). The study determined that the developed application could perform its intended functions and has met the ISO ISO/IEC 25010 Standards Evaluation Metric.



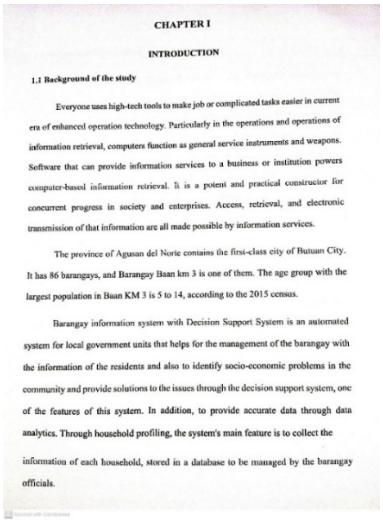
ABSTRACT

The implementation of knowledge discovery, commonly referred to as data mining, has been considered a latent solution for the containment of the COVID-19 pandemic. The proponents developed a prototype symptom-monitoring web application as a helping tool for the concerned individuals who may have been exposed to COVID-19. By integrating Apriori Algorithm and Naive Bayes Classifier; Apriori Algorithm to mine association rules to predict the following occurring symptom based on confidence percentage, and Naïve Bayes Classifier to analyze and predict the individual's risk exposure based on her/his details, the prototype system discovered that the highest risk exposure with 64% was if the individual was working on the frontline (frontliner) and the symptoms such as fever, cough and sore throat are the common sequence of indication with 56% confidence of risk exposure. The system aimed to address the alarming concern of being exposed to the virus due to manual monitoring, offer a platform for the individuals to track their symptoms within the entire seven-day monitoring course and produce data-driven predictions to give insights about the risk level of COVID-19 exposure. The proponents believe that with intensive research using sufficient data and the right tools for development, the prototype can become a potential tool to aid the community and healthcare organizations.

ABSTRACT

The implementation of knowledge discovery, commonly referred to as data mining, has been considered a latent solution for the containment of the COVID-19 pandemic. The proponents developed a prototype symptom-monitoring web application as a helping tool for the concerned individuals who may have been exposed to COVID-19. By integrating Apriori Algorithm and Naive Bayes Classifier; Apriori Algorithm to mine association rules to predict the following occurring symptom based on confidence percentage, and Naive Bayes Classifier to analyze and predict the individual's risk exposure based on discovered that the highest risk exposure her/his details, the prototype system with 64% was if the individual was working on the frontline (frontliner) and the hroat are the common sequence of symptoms such as fever, cough and sore t indication with 56% confidence of risk exposure. The system aimed to address the alarming concern of being exposed to the virus due to manual monitoring, offer a platform for the individuals to track their symptoms within the entire seven-day monitoring course and produce data-driven predictions to give insights about the risk level of COVID-19 exposure. The proponents believe that with intensive research using sufficient data and the right tools for development, the prototype can become a potential tool to aid the community and healthcare organizations.

Fig. 2. Captured text image and digital text extracted thru OCR result from abstract (left-to-right image)



CHAPTER I

INTRODUCTION

1.1 Background of the study

Everyone uses high-tech tools to make job or complicated tasks easier in current era of enhanced operation technology. Particularly in the operations and operations of information retrieval, computers function as general service instruments and weapons. Software that can provide information services to a business or institution powers computer-based information reirievai. Ii is a potent and practical constructor for concurrent progress in society and enterprises. Access, retrieval, and electronic transmission of that information are all made possible by information services. The province of Agusan del Norte contains the first-class city of Butuan City. It has 86 barangays, and Barangay Baan km 3 is one of them. The age group with the largest population in Baan KM 3 is 5 to 14, according to the 2015 census. Barangay information system with Decision Support System is an automated system for local government units that helps for the management of the barangay with the information of the residents and also to identify socio-economic problems in the community and provide solutions to the issues through the decision support system, one of the features of this system. In addition, to provide accurate data through data analytics. Through household profiling, the system's main feature is to collect the information of each household, stored in a database to be managed by the barangay officials.

CHAPTER I

INTRODUCTION

1.1 Background of the study

Everyone uses high-tech tools to make job or complicated tasks easier in current era of enhanced operation technology. Particularly in the operations and operations of information retrieval, computers function as general service instruments and weapons. Software that can provide information services to a business or institution powers computer-based information reirievai. Ii is a potent and practical constructor for concurrent progress in society and enterprises. Access, retrieval, and electronic transmission of that information are all made possible by information services. The province of Agusan del Norte contains the first-class city of Butuan City. It has 86 barangays, and Barangay Baan km 3 is one of them. The age group with the largest population in Baan KM 3 is 5 to 14, according to the 2015 census. Barangay information system with Decision Support System is an automated system for local government units that helps for the management of the barangay with
the information of the residents and also to identify socio-economic problems in the community and provide solutions to the issues through the decision support system, one of the features of this system. In addition, to provide accurate data through data analytics. Through household profiling, the system's main feature is to collect the information of each household, stored in a database to be managed by the barangay officials.

Fig. 3. Captured text image and digital text extracted thru OCR result from BOS (left-to-right image)

Stemmed

['abstract', 'abund', 'hotel', 'recommend', 'system', 'avail', 'internet', 'trivago', 'grab', 'hostel', 'zillow', 'mobil', 'problem', 'aris', 'find', 'nearbi', 'hous', 'rental', 'butuan', 'citi', 'implement', 'hous', 'rental', 'manag', 'system', 'develop', 'provid', 'optim', 'solut', 'physic', 'find', 'rental', 'hous', 'paperbas', 'process', 'manag', 'rental', 'hous', 'system', 'find', 'nearbi', 'hous', 'rental', 'avail', 'make', 'comparison', 'rental', 'hous', 'tenant', 'make', 'choic', 'accord', 'featur', 'requir', 'method', 'data', 'collect', 'collect', 'user', 'requir', 'interview', 'sourc', 'data', 'system', 'function', 'primari', 'data', 'develop', 'system', 'follow', 'waterfal', 'model', 'methodolog', 'system', 'interfac', 'written', 'design', 'use', 'python', 'clientsid', 'script', 'use', 'javascript', 'serversid', 'script', 'use', 'python', 'system', 'three', 'user', 'administr', 'owner', 'tenant', 'propon', 'care', 'analyz', 'necessari', 'data', 'gather', 'internet', 'conduct', 'survey', 'assess', 'qualiti', 'system', 'thirti', 'respond', 'compos', 'hous', 'rental', 'owner', 'twenti', 'hous', 'renter', 'assess', 'accumul', 'total', 'mean', 'equival', 'highli', 'accept', 'devic', 'inde', 'ideal', 'solut', 'help', 'landlord', 'effici', 'manag', 'inform', 'digit', 'return', 'make', 'easi', 'custom', 'search', 'housestolet', 'sinc', 'help', 'landlord', 'advertis', 'rental', 'hous', 'keyword', 'hous', 'rental', 'manag', 'system', 'softwar', 'python', 'javascript', 'housetolet']

Stemmed

['chapter', 'introduct', 'background', 'studi', 'everyon', 'use', 'hightech', 'tool', 'make', 'complic', 'task', 'easier', 'current', 'enhanc', 'oper', 'technolog', 'particularli', 'oper', 'oper', 'inform', 'retriev', 'comput', 'function', 'gener', 'servic', 'instrument', 'weapon', 'softwar', 'provid', 'inform', 'servic', 'busi', 'institut', 'power', 'computerbas', 'inform', 'reirieval', 'potent', 'practic', 'constructor', 'concurr', 'progress', 'societi', 'enterpris', 'access', 'retriev', 'electron', 'transmiss', 'inform', 'made', 'possibl', 'inform', 'servic', 'provinc', 'agusan', 'nort', 'contain', 'firstclass', 'citi', 'butuan', 'citi', 'barangay', 'barangay', 'baan', 'group', 'largest', 'popul', 'baan', 'accord', '2015', 'censu', 'barangay', 'inform', 'system', 'decis', 'support', 'system', 'autom', 'system', 'local', 'govern', 'unit', 'help', 'manag', 'barangay', 'inform', 'resid', 'identifi', 'socioeconom', 'problem', 'commun', 'provid', 'solut', 'issu', 'decis', 'support', 'system', 'featur', 'system', 'addit', 'provid', 'accur', 'data', 'data', 'analyt', 'household', 'profil', 'system', 'main', 'featur', 'collect', 'inform', 'household', 'store', 'databas', 'manag', 'barangay', 'offici']

Fig. 4. Pre-processed and digitized text of abstract and BOS (left-to-right image)

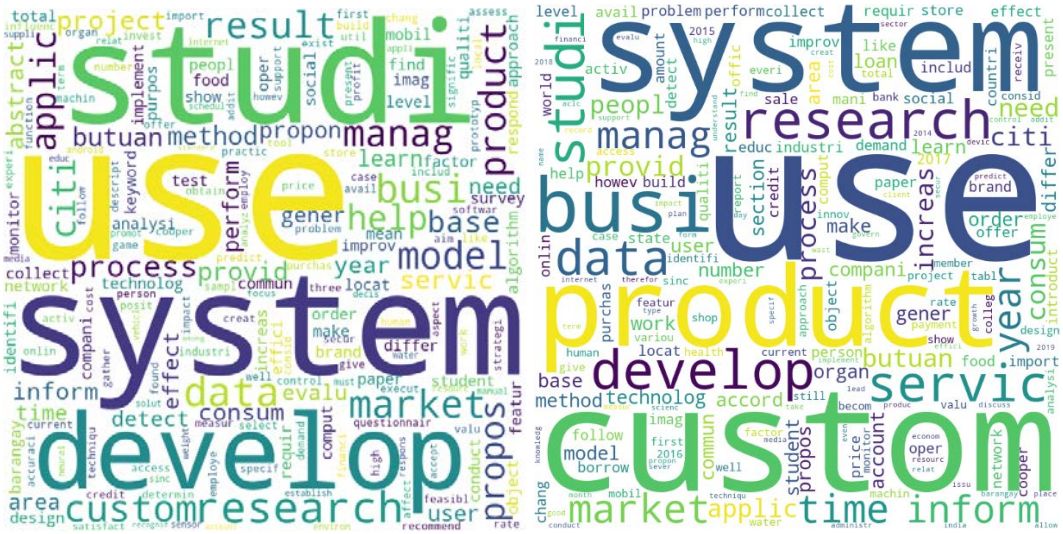


Fig. 5. Word cloud term frequency of abstract and BOS (left-to-right image)

TABLE I. TERM FREQUENCY DISTRIBUTION

Dataset	Term	Capstone	Thesis	Case Study	Feasibility Study	Total
Abstract	use	730	579	438	172	1919
	system	1246	415	19	30	1710
	studi	371	269	719	272	1631
	develop	680	163	135	86	1064
	research	190	188	471	140	989
Background of Study	use	657	1290	413	194	2554
	custom	128	142	2051	188	2509
	system	1278	859	128	92	2357
	product	140	137	1177	660	2114
	busi	155	100	869	828	1952

TABLE II. MONO GLOBAL WEIGHTS FOR ABSTRACT AND BOS

Term	Abstract	Term	BOS
use	1.015216258	*use*	1.020203602
system	1.029146199	*system*	1.021688587
studi	1.017633354	*custom*	1.032698286
develop	1.02556391	*product*	1.022270577
research	1.019049545	*busi*	1.017807377

TABLE III. EVALUATION RESULTS USING ABSTRACT DATASET

Supervised TWS with ML Models	Precision	Recall	F1 Score
TF MONO + KNeighborsClassifier	0.797330445	0.752194575	0.7299703264
TF MONO + MultinomialNB	0.576546545	0.635416668	0.703264095
TF MONO + LinearSVC	0.903361503	0.860701865	0.884272997
SQRT TF MONO + KNeighborsClassifier	0.877848725	0.844245163	0.8635014837
SQRT TF MONO + MultinomialNB	0.910247093	0.86169793	0.8872403561
SQRT TF MONO + LinearSVC	0.943446485	0.93636149	0.940652819

TABLE IV. EVALUATION RESULTS USING BOS DATASET

Supervised TWS with ML Models	Precision	Recall	F1 Score
TF MONO + KNeighborsClassifier	0.781659458	0.781659458	0.7715133531
TF MONO + MultinomialNB	0.758196003	0.758196003	0.6617210682
TF MONO + LinearSVC	0.564104215	0.564104215	0.8308605341
SQRT TF MONO + KNeighborsClassifier	0.597599638	0.597599638	0.8664688427
SQRT TF MONO + MultinomialNB	0.866617705	0.866617705	0.8902077151
SQRT TF MONO + LinearSVC	0.79807299	0.79807299	0.940652819

TABLE V. ISO/IEC 25010 STANDARDS SURVEY RESULTS FOR RECAPP

ISO Characteristics	Ratings in Percentage (%)				
	5	4	3	2	1
Functional Sustainability	60%	36%	3%	1%	0%
Performance Efficiency	53.33%	40%	5.33	1.33%	0%
Usability	57%	39%	2%	2%	0%
Reliability	60%	36%	3%	1%	0%

VI. CONCLUSION

The researchers were able to develop the classification model and application to eliminate the inconvenience and lessen the time consumption on classifying research documents for Students and Instructors. To identify which supervised term weighting scheme and machine learning algorithm can be best paired considering the evaluation metrics of each combination, the researchers used state-of-the-art supervised term weighting schemes and machine learning algorithms to simulate a classification on gathered dataset. As proven by the experiment results, the combination of SQRT TF-MONO and Linear SVC has the highest precision and recall values, and most importantly, F1 Scores of 0.94 both for the abstract and the background of the study datasets and, therefore, should be used as the classification model to classify research documents. Moreover, the researchers have developed an application prototype where users can import and classify research papers in bulk using the developed classification model. An ISO/IEC 25010 standards survey is conducted, and according to the results, most of the respondents have responded positively.

Finally, the researchers have concluded that the application can be helpful for research advisers, panelists, and reviewers to speed up the classification time by categorizing multiple research papers at once instead of reading and manually analyzing them. Lastly, the researchers believe that the study can be improved in the future.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Mr. Junell T. Bojocan of the Computer Education Department for his invaluable support in the BS Computer Science program, as well as to Mr. Gabriel Adolfo C. Malbas of the Research Extension and Innovation Department of ACLC College of Butuan for his generous financial support. The authors recognize the instrumental role of their contributions in completing this study.

REFERENCES

[1] X. Luo, “Efficient English text classification using selected Machine Learning Techniques,” Alexandria Eng. J., vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.

[2] S.-W. Kim and J.-M. Gil, “Research paper classification systems based on TF-IDF and LDA schemes,” Human-centric Comput. Inf. Sci., vol. 9, no. 1, p. 30, 2019, doi: 10.1186/s13673-019-0192-7.

[3] D. Sarkar, Text Analytics with Python - A Practitioner's Guide to Natural Language Processing. 2019.

[4] A. Allahverdipour and F. S. Gharehchopogh, "An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification," J. Adv. Comput. Res., vol. 9, pp. 37–48, 2018.

[5] G. B. Holanda et al., "Development of OCR system on android platforms to aid reading with a refreshable braille display in real time," Meas. J. Int. Meas. Confed., vol. 120, pp. 150–168, 2018, doi: 10.1016/j.measurement.2018.02.021.

[6] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," Int. J. Med. Inform., vol. 144, no. June, p. 104302, 2020, doi: 10.1016/j.ijmedinf.2020.104302.

[7] C. C. Abalorio and M. Cerna, "Course Evaluation Generator (Ceg): An Automated Academic Advising System with Optical Character Recognition," Int. J. Technol. Eng. Stud., vol. 4, no. 5, pp. 189–196, 2018, doi: 10.20469/ijtes.4.10003-5.

[8] I. Alsmadi, "Term weighting scheme for short-text classification: Twitter corpuses," Neural Comput. Appl., vol. 8, 2018, doi: 10.1007/s00521-017-3298-8.

[9] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," Futur. Gener. Comput. Syst., vol. 101, pp. 341–371, 2019, doi: 10.1016/j.future.2019.06.022.

[10] X. Zhou et al., "A survey on text classification and its applications," Web Intell., vol. 18, no. 3, pp. 205–216, 2020, doi: 10.3233/WEB-200442.

[11] Q. Li et al., "A Survey on Text Classification: From Shallow to Deep Learning," arXiv, 2020, doi: 10.48550/ARXIV.2008.00364.

[12] Muna Ahmed Awel and A. I. Abidi, "Review on Optical Character Recognition," Int. Res. J. Eng. Technol., vol. 6, no. 6, pp. 3666–3669, 2019, [Online]. Available: www.irjet.net

[13] V. Z. V Singco, J. C. Trillo, C. C. Abalorio, J. C. M. Bustillo, J. T. Bojocan, and M. C. Elape, "OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with FinetuneTransformer Modelsfor Long Document," vol. 13, no. 02, 2023, doi: 10.46338/ijetae0223.

[14] P. Divya et al., "Web based optical character recognition application using flask and tesseract," Mater. Today Proc., 2021, doi: 10.1016/j.matpr.2020.10.850.

[15] T. Dogan and A. K. Uysal, "On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification," Arab. J. Sci. Eng., 2019, doi: 10.1007/s13369-019-03920-9.

[16] Z. Tang, W. Li, and Y. Li, "An improved supervised term weighting scheme for text representation and classification," Expert Syst. Appl., vol. 189, p. 115985, 2022, doi: https://doi.org/10.1016/j.eswa.2021.115985.

[17] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," J. Informetr., vol. 14, no. 4, p. 101076, 2020, doi: 10.1016/j.knosys.2012.06.005.

[18] C. C. Abalorio, R. P. Medina, A. M. Sison, and G. A. Dalaorao, "Extended Max-Occurrence with Normalized Non-Occurrence as MONO Term Weighting Modification to Improve Text Classification," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 91–97, 2022, doi: 10.14569/IJACSA.2022.0130411.

[19] C. C. Abalorio, A. M. Sison, R. P. Medina, and G. A. Dalaorao, "Applying EMONO Variants to Multi-Class Sentiment Analysis for Short-Distance Inter-Class Frequency of Term," vol. 71, no. 4, pp. 1938–1947, 2022.

[20] A. Basarkar, "Document Classification using Machine Learning," 2017. doi: https://doi.org/10.31979/etd.6jmu-9xdt.

[21] A. Barua, O. Sharif, and M. M. Hoque, "Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation," Procedia Comput. Sci., vol. 193, pp. 112–121, 2021, doi: https://doi.org/10.1016/j.procs.2021.11.002.

[22] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," PLoS One, vol. 15, no. 5, pp. 1–22, 2020, doi: 10.1371/journal.pone.0232525.

[23] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," Artif. Intell. Rev., vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.

[24] J. C. M. Bustillo, R. P. Medina, A. M. Sison and M. Y. Orong, "Predictive Hybridization Model integrating Modified Genetic Algorithm (MGA) and C4.5," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1500-1507, doi:10.1109/ICECA55336.2022.10009532.