EUL: A Digital Research Repository System

A Thesis Presented to

The Faculty of the School of Computer Studies

Department of the

University of San Jose-Recoletos

Cebu City, Philippines

In Partial Fulfillment

Of the Requirements for Thesis 1

Members

Cristopher Bohol

Paul Joshua Premacio

Thesis Adviser

Dr. Lorna Miro

Date 2023

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This is a test abstract. Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

CHAPTER I

INTRODUCTION

RATIONALE OF THE STUDY

Classifying is a process of assigning something or someone into different classes or categories based on shared quality or characteristics. Humans have done this process even before the invention of modern computers, Medieval army commanders sorting their formation based on the roles of the unit, pikes up-front, archers at the back, or a renaissance doctor categorizing medicines on a shelf and labeling them. The goal is simple: to easily manage and analyze the information at hand. A simple task no doubt, but doing it manually, with a large amount of information, let's say documents, well that's where the fun stops. Managing documents manually, or in this study's case, research projects, is not ideal, especially if we're manually categorizing hundreds of research from the repository. To improve this process, we enlist the help of machines in the form of Artificial Intelligence (AI).

The advancement of modern computing gave birth to AI and eventually its underlying fields, Machine Learning (ML) and Natural Language Processing (NLP). AI technology has been around since the 1940s [1]. It's been fine-tuned throughout the past decades. There are a lot of real-world applications that use ML and NLP. For instance: chatbots, language translators, email classification and filtering. With the help of AI and its related fields, document classification can now be done automatically [2].

In 2015, all United Nation member states adopted a resolution that calls for a shared blueprint for peace and prosperity, they call this, 'The 2030 Agenda for Sustainable Development. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership. With this plan in mind, classifying research with the UN SDGs makes the research presently relevant.

This study aims to develop a system that an educational institution can benefit from by creating a web and mobile application that serves as a repository of all research. In addition, research after uploading to the repository is automatically classified according to its SDGs and its research topics (Artificial Intelligence, Machine Learning, Computer Vision, etc.).

THEORETICAL BACKGROUND

Machine Learning

ML is defined as a type of AI that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so [3]. This can be done by using algorithms or models to analyze and draw inferences from patterns in the data. There are three types of Machine learning: *supervised*, *unsupervised*, and *reinforcement learning*.

Supervised Learning is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labeled for a particular output. The model is trained until it can detect the underlying patterns and relationships between the input data and the output labels, enabling it to yield accurate labeling results when presented with never-before-seen data [4]. It can be further down into two categories, ***classification*** and ***regression*** algorithms.

A **classification** algorithm is defined as a technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups [5]. Examples of classification are SVM and Naive Bayes.

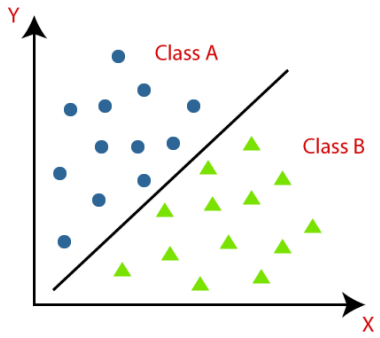


Figure 1. Sample Visualization of a Classification Algorithm

A **regression** algorithm is defined as a model that predicts the output values based on input features from the data fed into the system. Linear and Logistical regression are the most popular regression algorithms.

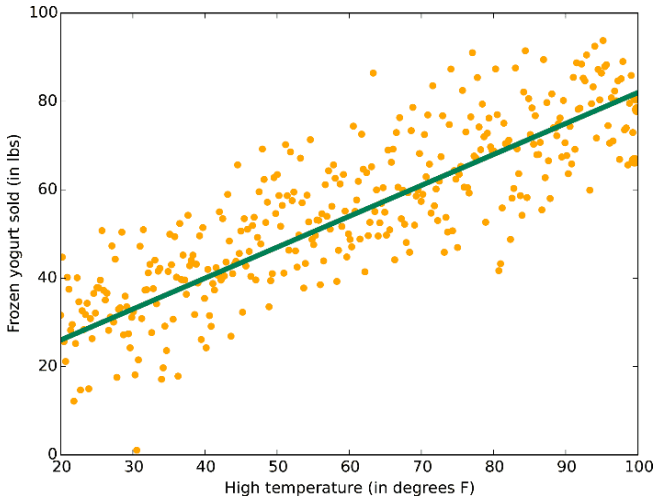


Figure 2. Sample Visualization of a Regression Algorithm

Support Vector Machines is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

Naive Bayes is a simple learning algorithm that utilizes the Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class [7]. Naive Bayes is simple and easy to implement as it does not require a large amount of data for it to accurately predict. One of the disadvantages of this algorithm is, that it treats all predictors as independent variables thus limiting the algorithm's usability in real-world scenarios.

K-Nearest Neighbor Algorithm also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

Unsupervised Learning refers to the use of AI algorithms to identify patterns in data sets containing data points that are neither classified nor labeled [6].

K-means Clustering is one of the simplest and most popular unsupervised machine learning algorithms. The objective of K-means is simple: group similar data points together and discover underlying patterns [8].

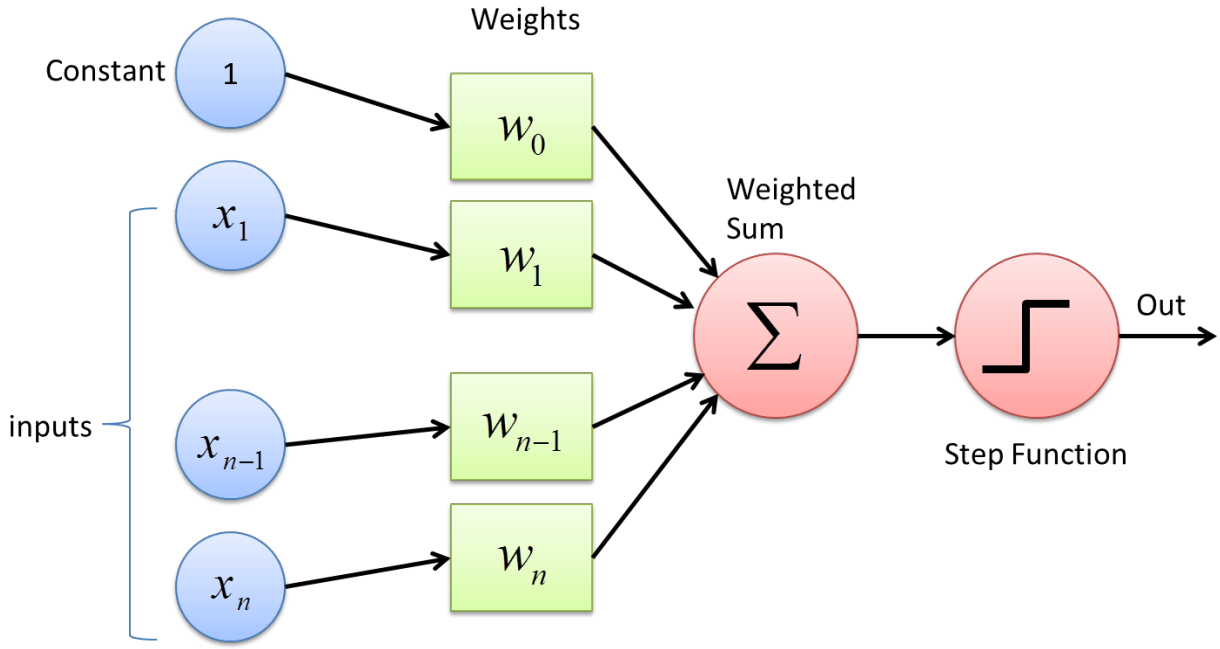


Figure 3. Sample Visualization of a Convolutional Neural Network

Convolutional Neural Network is a class of deep, feed-forward artificial neural networks (where connections between nodes do *not* form a cycle) & uses a variation of multilayer perceptrons designed to require minimal preprocessing. These are inspired by the animal's visual cortex.

Term Frequency - Inverse Document Frequency (TF - IDF)

TF-IDF is an example of text vectorization. It counts the frequency of a word in a document to measure its relevance to a document in a collection of documents. Text Vectorization is the process of converting text into numerical representation [9].

Term Frequency (TF) is simply just the count of a word in a document.

Inverse Document Frequency (IDF) is the frequency of the word across all documents. This gets the rarity of the word. The closer it is to 0, the more common the word is. Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document [10].

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Figure 4. TF-IDF sample

Mathematically it can be described as,

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ idf\ (t, D)$$

Where,

$$tf\ (t, d) = log\ (1 + freq\ (t, d))$$

$$idf\ (t, D) = log\ (\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

Natural Language Processing

Computers can't understand human language. It only understands 0s and 1s or binary information. For computers and humans to *communicate directly,* NLP is used. NLP is described as a branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can [11].

REVIEW OF THE RELATED LITERATURE

Research paper classification systems based on TF-IDF and LDA schemes

A research paper classifier that uses TF-IDF as its text vectorization algorithm and LDA scheme in topic modeling [13]. This study uses the K-means clustering algorithm to finally classify its test data. While there are similarities between both studies, the study in question only used the abstract part of the research as its source of data, EUL on the other hand, uses the abstract, the rationale of the study, and the research methodology as its inputs.

Subject classification of research papers based on interrelationships analysis

A similar study of classifying research papers uses a novel supervised approach for subject classification of scientific articles based on an analysis of their interrelationships [14]. The study in question uses citations, common authors, and common references to assign subjects to papers. The study focuses more on scientific articles whereas EUL accepts any form of the research topic as long as its authors are affiliated with USJ-R.

PROJECT OBJECTIVES

This study intends to develop a mobile and web application for managing USJ-R's research repository. Specifically, this study aims to

- To have a single research repository for the entire USJ-R community.
- To classify projects based on the UN's Sustainable Development Goals (SDG) and different research topics.

PROJECT SCOPE AND LIMITATIONS

This study outputs a web and mobile application. The web component handles the processing of the algorithms as well as UI interaction. The mobile component, on the other hand, has features similar to the web component excluding the processing of the algorithms. The mobile component requires Android v4.1 (Jellybean) and above. The system only supports English as its main language. As of now, the application will only be used by USJR faculty and students.

SIGNIFICANCE OF THE STUDY

Presently, there is a lack of an institutional-wide research repository system. The study not only is a research repository but also a system that automatically classifies research in accordance with the United Nation's Sustainable Development Goals (UN SDG) as well as different research topics (AI, ML, Computer Vision).

United Nation's 17 Sustainable Development Goals:

- No Poverty
- Zero Hunger
- Good Health and Well-Being
- Quality Education
- Gender Equality
- Clean Water and Sanitation
- Affordable and Clean Energy
- Decent Work and Economic Growth
- Industry, Innovation, and Infrastructure
- Reduced Inequalities
- Sustainable Cities and Communities
- Responsible Consumption and Production
- Climate Action
- Life Below Water
- Life on Land
- Peace, Justice, and Strong Institutions
- Partnerships for the Goals

RESEARCH METHODOLOGY

Classification

The study relies on the inputted research abstract, rationale of the study, and research methodology. The inputted data will then be used to analyze, classify and give proper recommendations to the users.

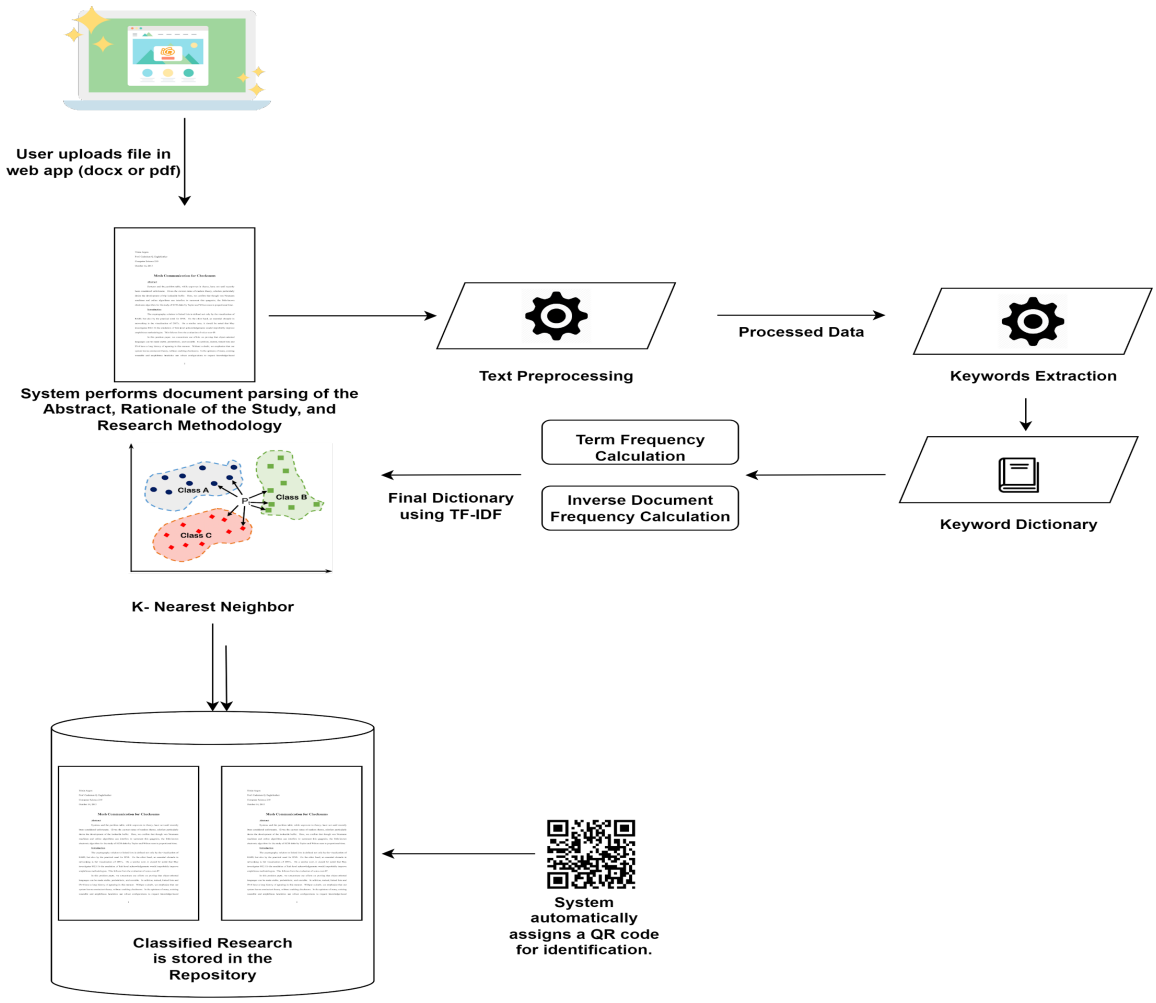


Figure 5 Conceptual Framework for Research Classification

The Abstract, the Rationale of the study, and the Research Methodology, were chosen to be used as the primary source of data since most of the relevant keywords needed for the system to classify can be found in these documents. The classified data can be used in recommending research and checking for the novelty of research proposals.

In the first step, the user uploads a file (docx or pdf), and the system will perform document parsing to extract texts out of the Abstract, the Rationale of the Study, and the Research Methodology.

The second step in the classification is text preprocessing, in this area, the chosen research parts are cleaned of unnecessary text, like stop words, and punctuations. Just like in any machine learning model, text preprocessing plays an integral part in the accuracy of the

model as it removes unwanted or unimportant texts. Regular Expression, Stemming, and Lemmatization techniques are used in this process.

Once preprocessing is done, words are assigned a numerical value, that value represents the frequency of the word in the document. This process is called text vectorization. The system creates a frequency dictionary of the preprocessed data.

After creating the dictionary, TF-IDF calculates the relevance of a word in the entire collection of documents. This process gives us an overview of the uniqueness of the word in the entire corpus.

The result of the TF-IDF calculation is fed into our machine learning model, which is the K-Nearest Neighbor. This algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

The data set that will be used are prepared manually, research papers in the internet are carefully chosen to be used as one of the samples for the data set. Overall, 51 research papers will be used and will be preproccesed manually like removing unnecessary pages, or pages that contains only images. These papers will be trimmed down to only pages that are deemeed relevant like the abstract, introduction, and research methodology. However the process of extracting and creating Term Frequency for each papers are done automatically.

CHAPTER II

SOFTWARE REQUIREMENTS AND DESIGN SPECIFICATIONS

This chapter specifies the user and system requirements that are expected to be accomplished as well as the structure and process of achieving these. It contains sections for the Use Case Diagram, Use Case Narrative, Activity Diagram, Class Diagram, and User Interface Design,

USE CASE DIAGRAM

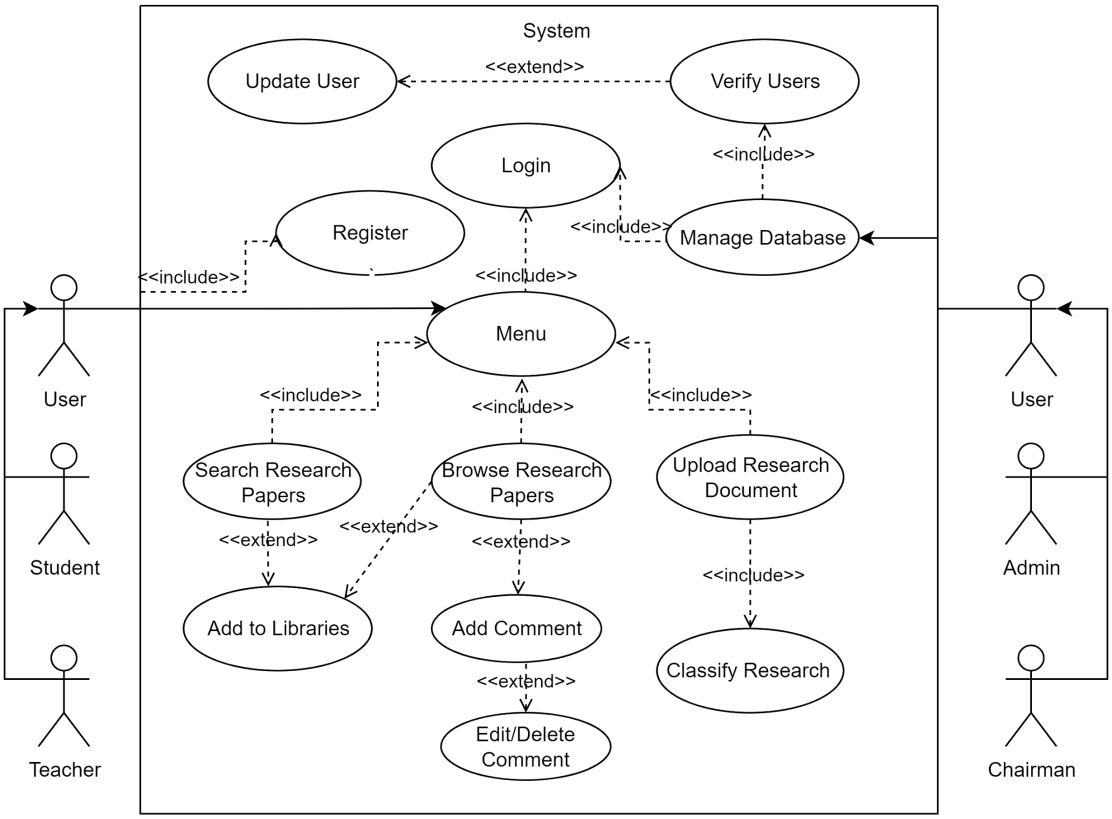


Figure 6 Defines the different use cases in which the system performs when student, teacher, and admin querying the research repository on the Application.

USE CASE NARRATIVE

The use case narrative shows the main scenarios and alternative flows of a use case. It provides more details about the use case.

Use Case 001	Verify User
Actor(s)	Admin, Chairman
Overview	Admin will send an api request to verify the user
Precondition(s)	User is not yet approved by the admin.
Post Condition(s)	The user is remove in the list of the pending user
Flow of Events	
Actor Action(s)	System Response
2. The admin will check if new request has been submitted on dashboard.	1. The system will display list of new user in the screen
3. The admin will approve the request of the user	
Alternate Flow of Events: A2. User is not affiliated	
1. If the submitted form from user is not recognizable by the chairman, the chairman will delete the user	
Alternate Flow of Events: A2. User info is incorrect	
1. If the input value is incorrect, the admin can alter the info provided by the user	

Use Case 002	Manage Database
Actor(s)	Admin
Overview	The admin will check and manage the database to avoid malicious attacks on the database where it securely stores private accounts of teachers and students and the submitted outputs of the Josenian Community.
Precondition(s)	The admin should be logged in already before managing the database.
Post Condition(s)	The app can connect and fetch data from the database.
Flow of Events	
Actor Action(s)	System Response
	1. Performs CRUD operations as requested by the admin.

Use Case 003	Add to Libraries
Actor(s)	Student, Teacher, System
Overview	Users can add research to their respective libraries.
Precondition(s)	Teachers or Students should be login with their respective accounts
Post Condition(s)	Selected research is saved in the library.
Flow of Events	

Actor Action(s)	System Response
	1. After clicking the add button, the research will be saved in their profile.

Use Case 004	Upload Research Document
Actor(s)	Admin, Student, System
Overview	Admin, and student can upload a soft copy of research to the database where it securely stores and classifies research according to what categories of Sustainable Development Goals the research should be classified.
Precondition(s)	Admin or student should be login to access the features upload button.
Post Condition(s)	Research has been successfully classified and stored in the research repository.
Flow of Events	
Actor Action(s)	System Response
	1. The user taps the button to upload research
	2. The system classifies the research.
	3. After classifying the research into its respective categories, the result screen would display the category result and the information of the research paper to be displayed on the front page such as author, adviser, and, etc.
Alternate Flow of Events: A1 Wrong Format	
	1. The system will display "Wrong Format! PDF only!".
	2. The user will pick another file until the user uploaded the PDF format.
	3. Continue Process UC4 num 2.

Use Case 005	Search Research
Actor(s)	Teacher, Student
Overview	To enable users to search the research repository.
Precondition(s)	Actors should be login with their respective accounts for them to access the search function
Post Condition(s)	none
Flow of Events	
Actor Action(s)	System Response
	1. Check if the user specifies categories or filters. 2. Check research in the repository based on user search.
	3. Displays a list of research based on the user's search.
Alternate Flow of Events: Alternate Flow A1: No research found.	
1. System displays text that no research is found.	

Use Case 006	Register
Actor(s)	Teacher, Student

Overview	Actors are expected to create their accounts before they can log in and use exclusive features of the app.
Precondition(s)	none
Post Condition(s)	none
Flow of Events	
Actor Action(s)	
	1. From home, the user will need to go to registration page
	2. The user will enter the necessary information and then click the register button
	3. The system will check if the entered email or id has already been registered before.
	3. If the account is not yet registered, the user is successfully registered to the website.
Alternative Flow of Events: A1: Exist Account	
If the system detects it is already existing, the system will pop up an error message	
The user will ask to input again.	

Use Case 007	Login
Actor(s)	Student, Teacher
Overview	To let the actors access private features in the app.
Precondition(s)	none
Post Condition(s)	none
Flow of Events	
Actor Action(s)	
	1. From the homepage, actors need to click sign in to redirect to the sign-in page.
	2. The actors will then input their account information like email and password.
	3. If the credential is valid and the account is approved, the actors will be redirected to the private homepage which is based on what role they have as a user status.
Alternate Flow of Events: A1: Credentials are not valid	
	1. The system will display the error message then the system will ask to input again.
	2. If the submitted credentials are correct, redirect to the private homepage.
	2. If the system detects that the number of an attempt made by the user is equal to 3, then the system will notify the user who wants to reset the password.
Alternate Flow of Events: A1: Credentials are valid but the account was not yet approved by admin.	
	1. System will display the error message to wait for a confirmation

	message.

Use Case 008	Classify Research
Actor(s)	System
Overview	It automatically classifies what kind of category the research belongs to.
Precondition(s)	The file uploaded has successfully extracted information
Post Condition(s)	Teacher successfully gave updates and added new instructions to the teams.
Flow of Events	
Actor Action(s)	System Response
	1. The system reads the text in the docs or pdf uploaded by the user.
	2. System performs preprocessing of the document.
	3. System performs the process indicated in the Classification Diagram.
	4. System successfully classified research.

Use Case 009	Add Comment
Actor(s)	Student, Teacher
Overview	This use case allows a student or teacher to add a comment to a research paper
Precondition(s)	
	The user is logged into the system.
	The user has selected a research paper to view.
	The research paper has comments enabled.
Post Condition(s)	
	The comment is added to the research paper in the system.
	The user is returned to the research paper view page.
Flow of Events	
Actor Action(s)	System Response
	1. The system displays a form for the user to enter their comment.
2. The user enters their comment and submits the form.	3. The system adds the comment to the research paper and displays it on the page.
	4. The user is returned to the research paper view page.
Alternate Flows	
2a. The user cancels the comment.	
1. The system discards the comment and returns the user to the research paper view page.	
2b. The user submits an empty comment.	
1. The system displays an error message indicating that the comment cannot be empty.	
2. The user is prompted to enter a valid comment.	
Exceptions:	
1. The user is not logged into the system.	

2. The user selects a research paper that does not have comments enabled.
3. The system encounters an error while adding the comment.

Use Case 010	Update User
Actor(s)	Admin, Chairman
Overview	This use case allows the admin to update a user's information in the library management system.
Precondition(s)	
	The actors is logged into the system.
	The actors has selected the "Update User" option from the dashboard panel.
	The actors to be updated has been selected.
Post Condition(s)	
	The user's information has been updated in the system.
Flow of Events	
Actor Action(s)	System Response
	1. The system displays a form with the current user information.
2. The user enters their comment and submits the form.	3. The system adds the comment to the research paper and displays it on the page.
	4. The user is returned to the research paper view page.
Alternate Flows	
3a. The admin cancels the update.	
1. The system discards the changes and returns the admin to the previous screen.	
Exceptions:	
1. The user is not logged into the system.	
2. The system encounters an error while updating the user information.	

Use Case 011	Browse Research Papers
Actor(s)	Admin, Chairman, Teachers, Students
Overview	This use case allows a user to browse the research papers in the library management system.
Precondition(s)	
	The actors is logged into the system.
	The actors has selected either one of the sdg icon in the home menu
Post Condition(s)	
	The user has viewed the selected research paper
Flow of Events	
Actor Action(s)	System Response

	1. The system displays a list of sdg in the home page
2. The user clicks one of the icon corresponds to the specific SDG category	3. The system will redirect the user to corresponding sdg where the sytem filter out the list of researches according to category
4. The users select read more to view the specific research	4. The user has viewd the selected research paper
Alternate Flows: none	
Exceptions:	
1. The user is not logged into the system.	
2. There are no research papers available in the system	

Use Case 012	Edit/Delete Coment
Actor(s)	Student, Teacher
Overview	This use case allows a user to edit or delete their comment on a research paper in the library management system.
Precondition(s)	
	The user is logged into the system.
	The user has selected a research paper to view.
	The user has previously added a comment to the research paper.
Post Condition(s)	
	The comment is added to the research paper in the system.
	The user is returned to the research paper view page.
Flow of Events	
Actor Action(s)	System Response
	1. The system displays the research paper information and the user's comment.
3. The user selects the "Edit/Delete Comment" option.	2. The system displays the edit or delete button.
4. The user edits or deletes their comment and submits the form.	5. The system updates or deletes the comment and displays a message indicating that it was updated or deleted successfully.
Alternate Flows: None	
Exceptions:	
1. The user is not logged into the system.	
2. The selected research paper does not exist.	
3. The user's comment does not exist.	
4. The user is not the author of the comment.	

Use Case 013	Search Research Papers
Actor(s)	Admin, Chairman, Teachers, Students
Overview	This use case allows a user to search for research papers in the library management system.
Precondition(s)	
	The user is logged into the system.
	The user has selected the "Search" option from the side navigation menu
Post Condition(s)	
	The comment is added to the research paper in the system.
	The user is returned to the research paper view page.
Flow of Events	
Actor Action(s)	System Response
	1. The system displays a search form with list of all researches
2. The user enters search keywords	3. The system displays a list of research papers that match the search criteria.
4. The user selects a research paper to view.	5. The system displays the selected research paper and its details.
Alternate Flows	
3a. The search returns no results.	
1. The system displays a message indicating that no results were found.	
Exceptions:	
1. The user is not logged into the system.	
2. There are no research papers available in the system.	
3. The search criteria are invalid.	

ACTIVITY DIAGRAM

The activity diagram shows the workflow behavior of the system by describing the sequence of actions in the process.

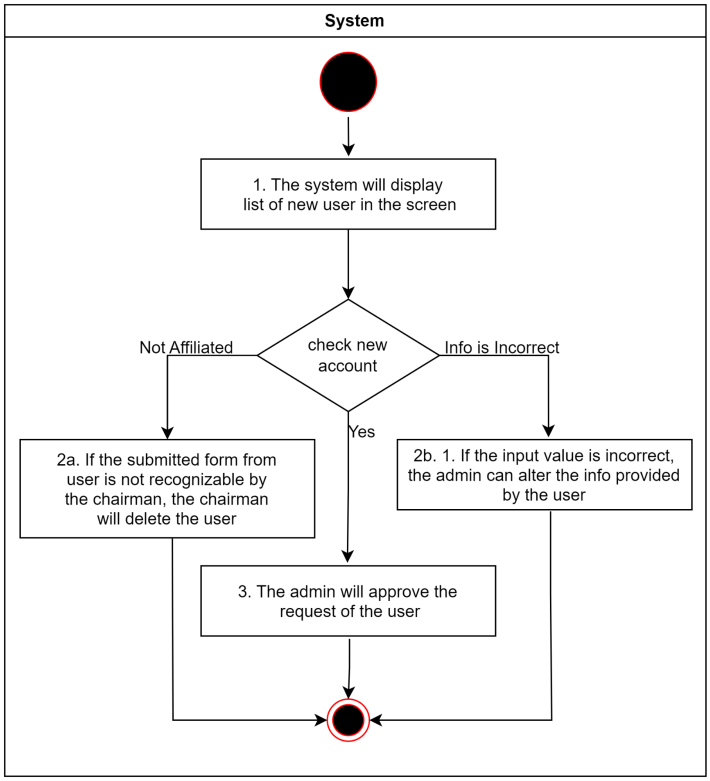


Figure 4. UC001 Verify User

Figure 4 shows the visualization when the admin verifies the request of the user. The system will check the inputted Student ID if it is correct, and the system will permit the user to access the app.

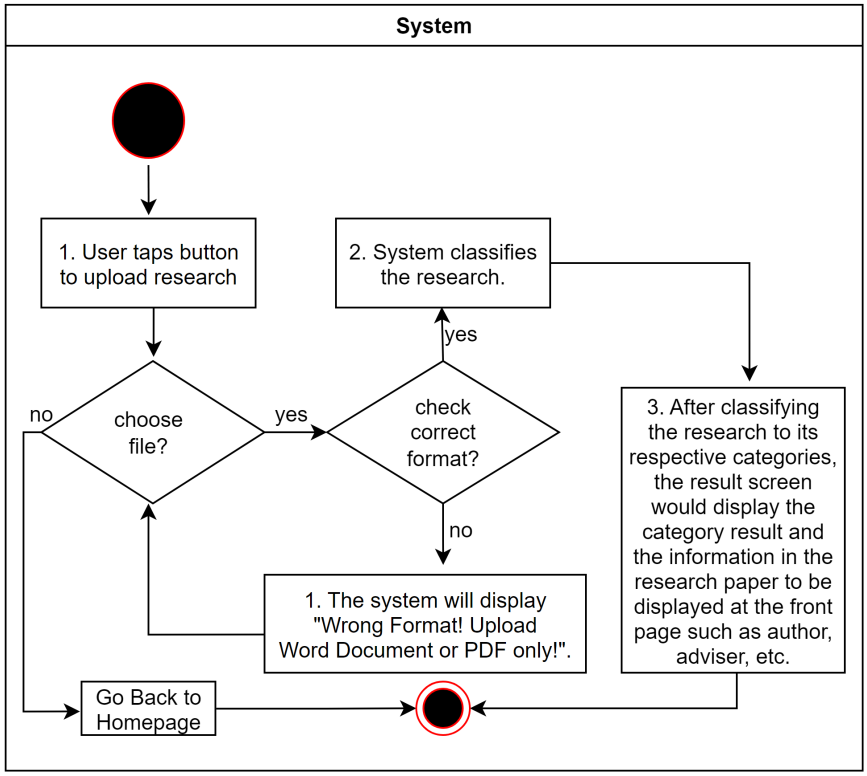


Figure 5. UC004 Upload Research Document

Figure 5 shows how the system behaves when the users upload research. The system will check the type of file being uploaded until it is a pdf file or a word document. The user can also cancel

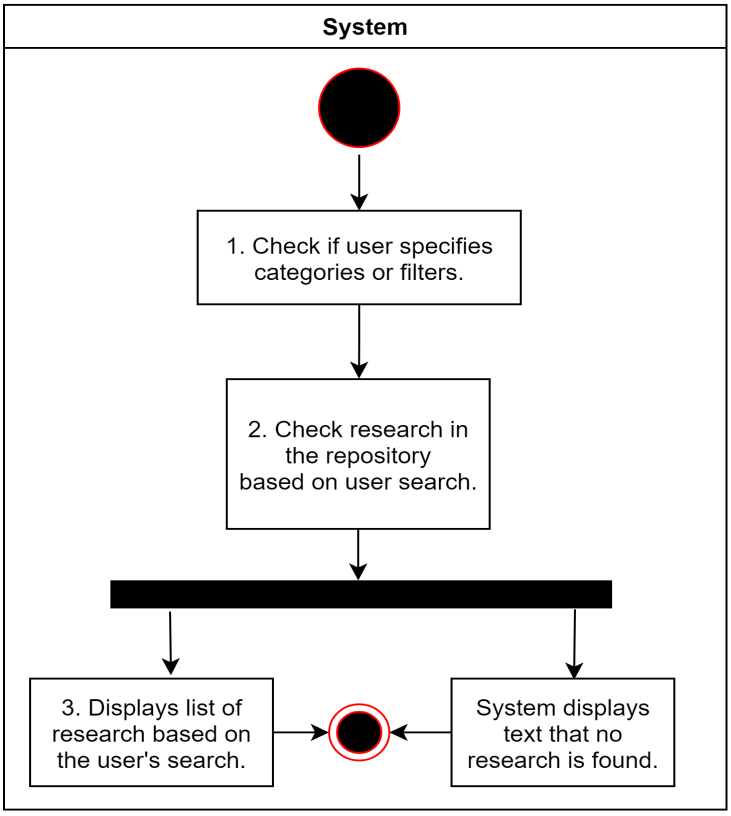


Figure 6. UC005 Search Research

Figure 6 shows the system's behavior when users either use filters with their search results or not. Research display is based on what categories are closer to the user input. It would only display no research found if stack categories plus the search term don't match up with the research in the repository.

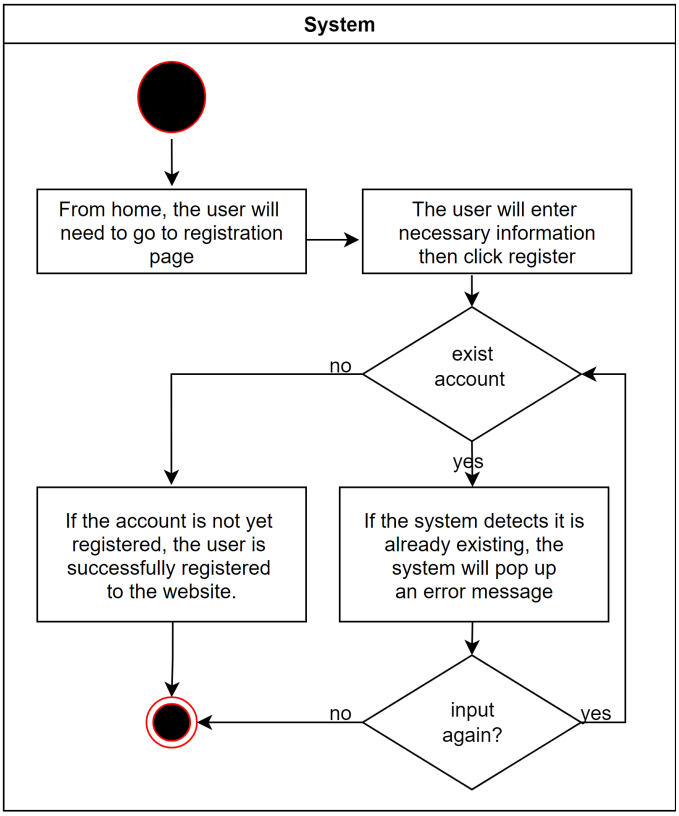


Figure 7. UC006 Register

Figure 7 shows the registration process of the user when using the application. If the account the user wants to register already exists then the user is required to enter again unless the user gives registers their account. Else, they will wait for the admin's approval of their account.

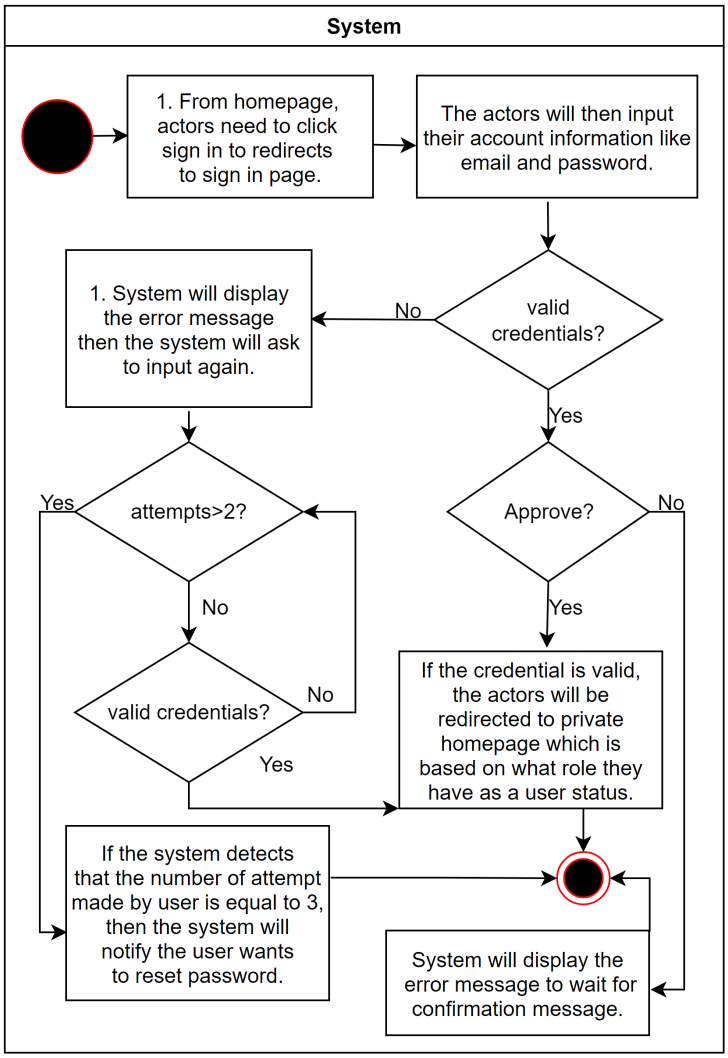


Figure 8. UC007 Login

Figure 8 shows the login process of the user when using the application. If the user fails to enter the correct information 3 times then the system will notify the user that he/she needs to reset the password using the user's email. Also, user's without approval from the admin may not be able to login.

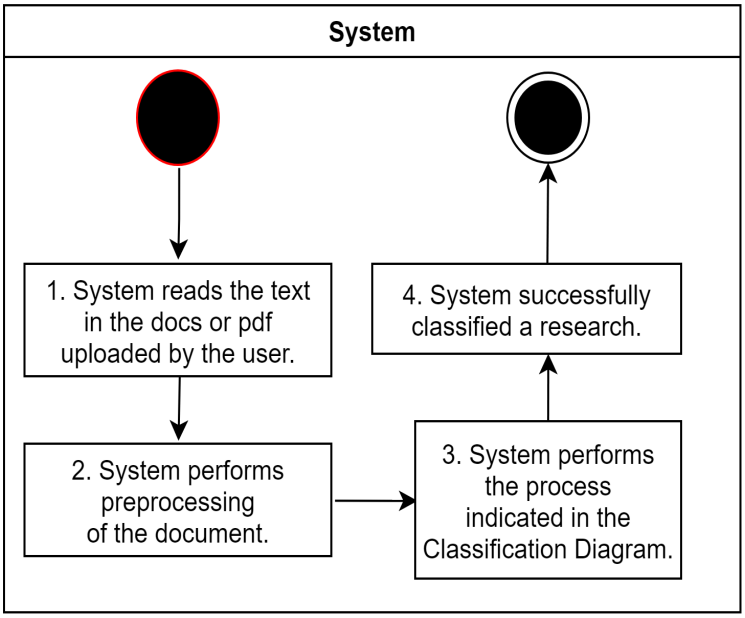


Figure 9. UC008 Classify Research

Figure 9 shows the behavior of the system when performing the classification of the research uploaded by the user. Extracted keywords from the Abstract and the rationale of the study will be used in order for the algorithm to be classified according to the predicted result. After classifying the research and, keywords, classified categories will be used in future use such as finding similar research.

CLASS DIAGRAM

The class diagram shows the structure of the system as classes with their attributes, operations, and relationships among other classes.

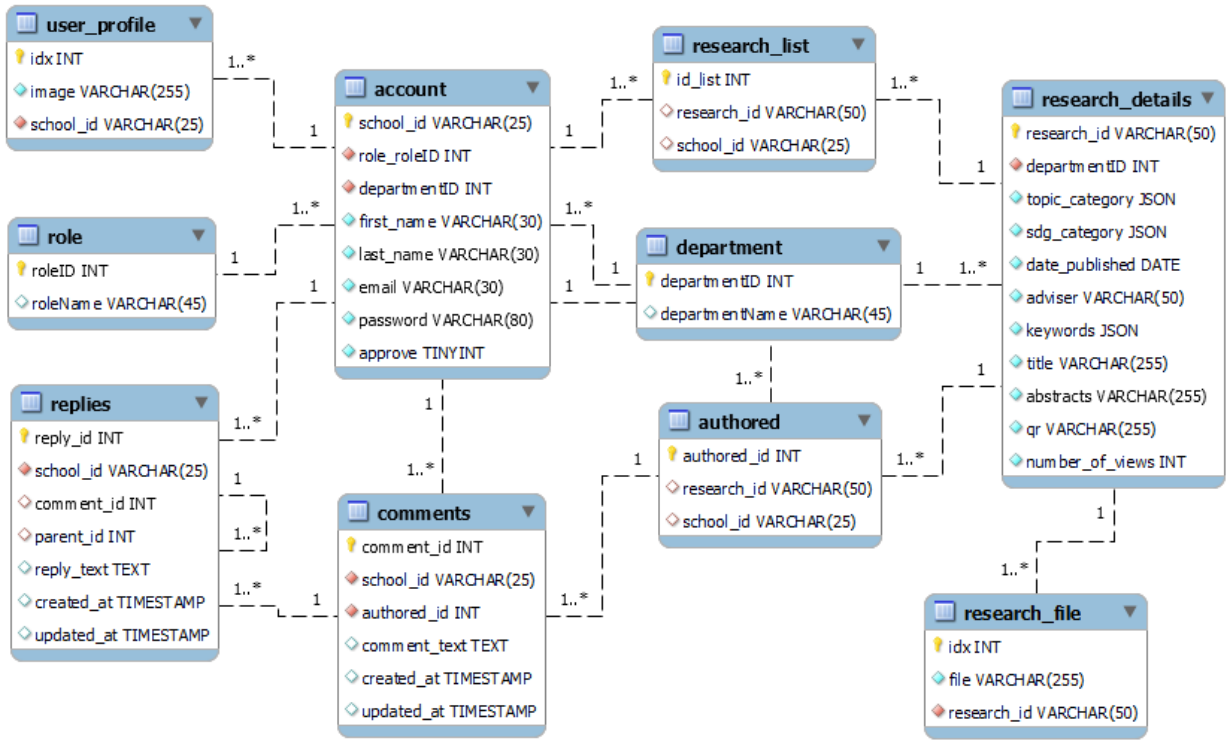


Figure 10. Class Diagram of the App System

USER INTERFACE DESIGN

The user interface design is the process of creating interfaces that are expected to provide insights into how the user can interact with the system.

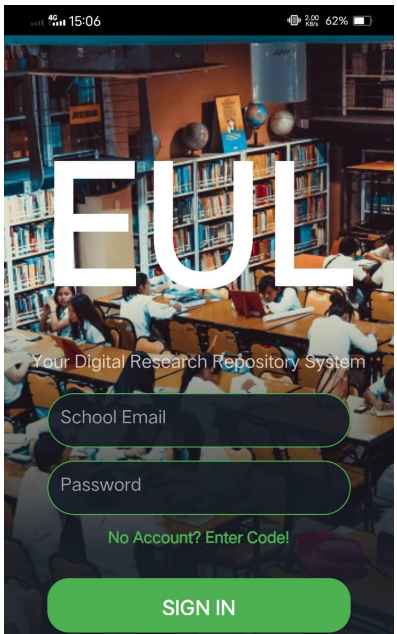
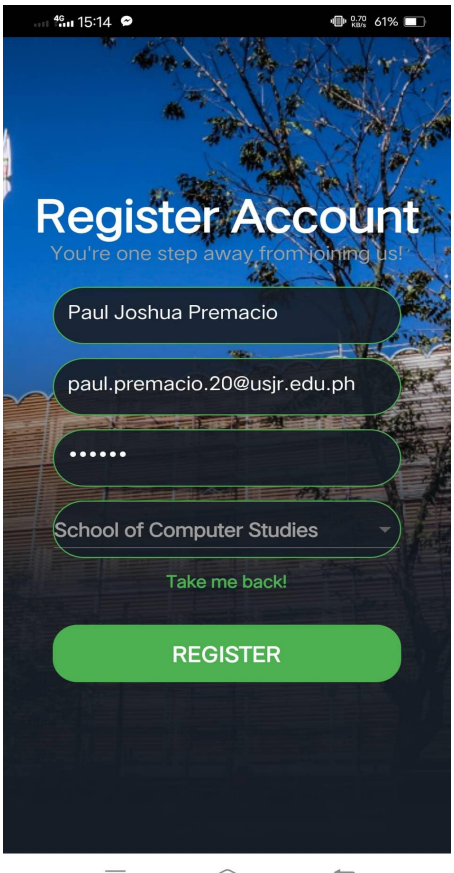


Figure 11. Login Screen

Figure 11 shows login where users can enter information to access the application.



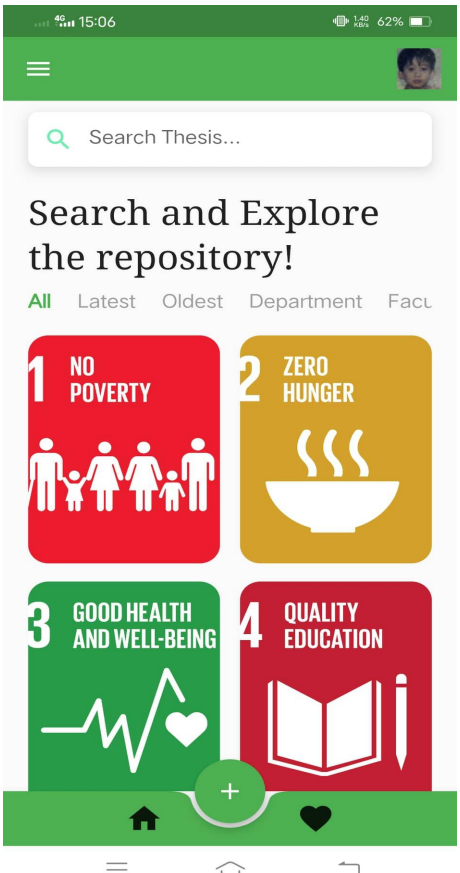


Figure 13. Register Screen *Figure 14. User Dashboard*

Figure 13 shows the register screen for the users with the required information. Figure 14 shows the user dashboard, where users can see different research based on different SDG goals. Students can also search for names of the research or sort it by category.

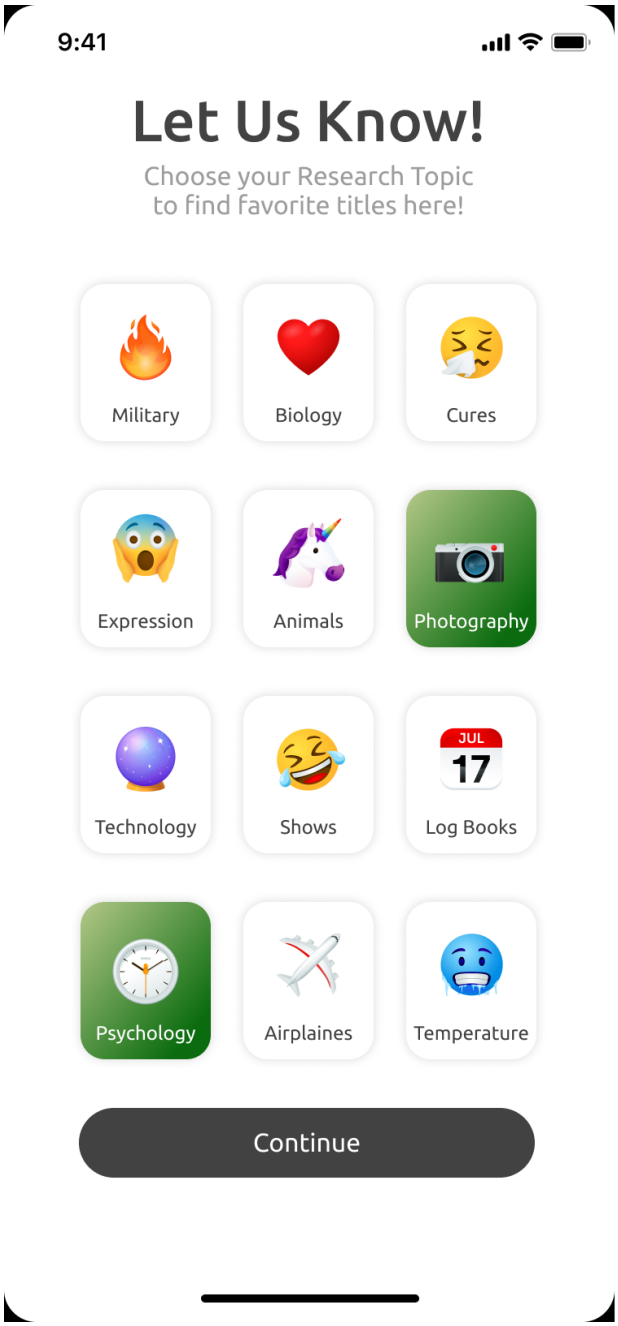


Figure 15. Likes Screen

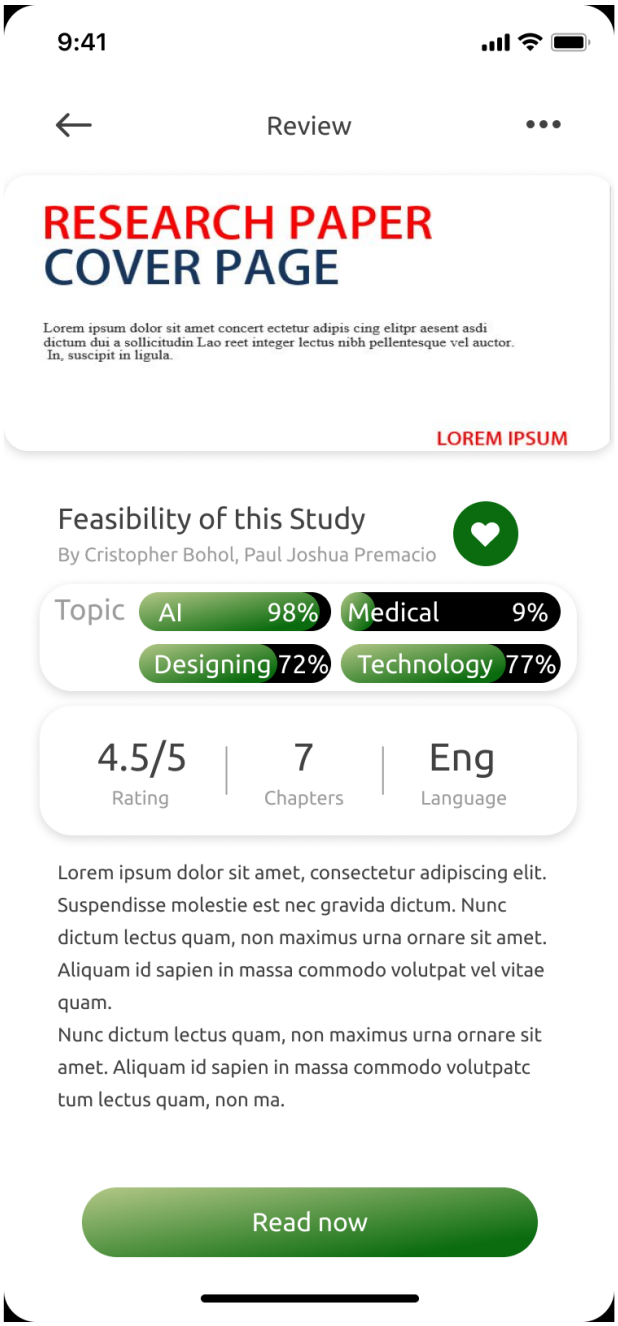


Figure 16. Research Screen

Figure 15 shows the screen after registration, this screen lets user picks whatever category they like. This is needed for the reference recommender system. Figure 16 shows the details of the research selected by the user.

Web User Interface

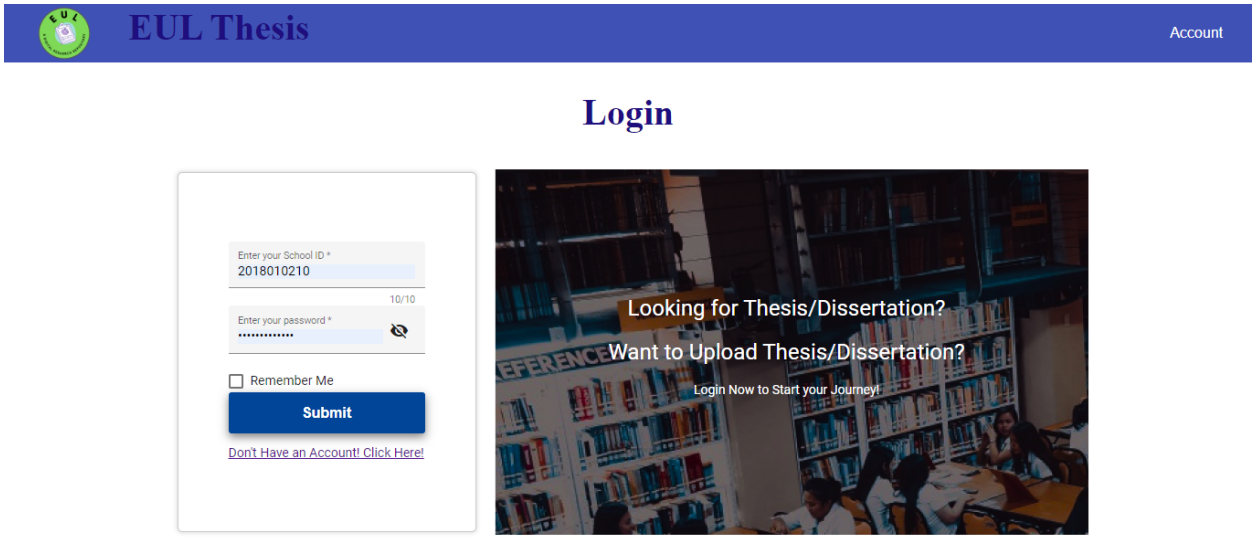


Figure 17. Web Homepage

Figure 17 shows the homepage of the web application. Here users can log in or be redirected to register an account.

Figure 18. Web User Dashboard

Figure 18 shows the user dashboard of the web application, users can view different papers that are categorized based on 17 different UN SDG goals.

EUL Thesis

Home | Cristopher Bohol | Logout

Cristopher Bohol

- Home
- Dashboard
- Search
- Library
- Logout

Chairman Dashboard

6	18	34
Total # of Chairman	Total # of Teacher	Total # of Student

58	75	9
Total # of Users	Total # of Research	Total # of Department

List of Student

Search

School ID	First Name	Last Name	Email	Role	Department	Approve	
0908603823	Ali	Hyland	pcvmbj959@nowhere.com	Student	School of Computer Studies	0	Edit Delete
1250949104	Abraham	Wallen	Petra.Wicker522@nowhere.com	Student	Senior High School	0	Edit Delete
1352014620	Cristobal	Cooke	Abbott@usjr.edu.com	Student	School of Engineering	0	Edit Delete
1522976795	Brigida	Foley	dzjqpz5688@usjr.edu.com	Student	Senior High School	0	Edit Delete
1587840093	Enrique	Sizemore	Rose_E.Shelby346@usjr.edu.com	Student	School of Allied Medical Sciences	0	Edit Delete

Items per page: 5 | 1 – 5 of 34

List of Teacher

Search

School ID	First Name	Last Name	Email	Role	Department	Approve	
0471594432	Ashley	Cook	ValVenegas559@usjr.edu.com	Teacher	School of Education	0	Edit Delete
0539869532	Nickie	Waite	Chatman@nowhere.com	Teacher	School of Education	1	Edit Delete
190changes	Harlan	Ludwig	LurleneMcnally@usjr.edu.com	Teacher	School of Allied Medical Sciences	0	Edit Delete
2020202020s	hdhf	hsdh	sdhs@gmail.com	Teacher	School of Law	0	Edit Delete
20666180210	Cristopher	Bohol	cris.bacus@gmail.com	Teacher	School of Business and Management	0	Edit Delete

Items per page: 5 | 1 – 5 of 18

EUL Thesis

Home Cristopher Bohol Logout

Cristopher Bohol

- Home
- Dashboard
- Search
- Library
- Logout

Go Back

Search

All | Department's Research | Teacher's Research | Student's Research

Unde voluptatem iste ullam eum.

Voluptas et perspiciatis. Non quod recusandae! Quia doloremque mollitia. Eveniet officia dignissimos! Recusandae hic natus. Omnis vitae consectetur. Numquam omnis iste; nihil deleniti debitis. Esse.

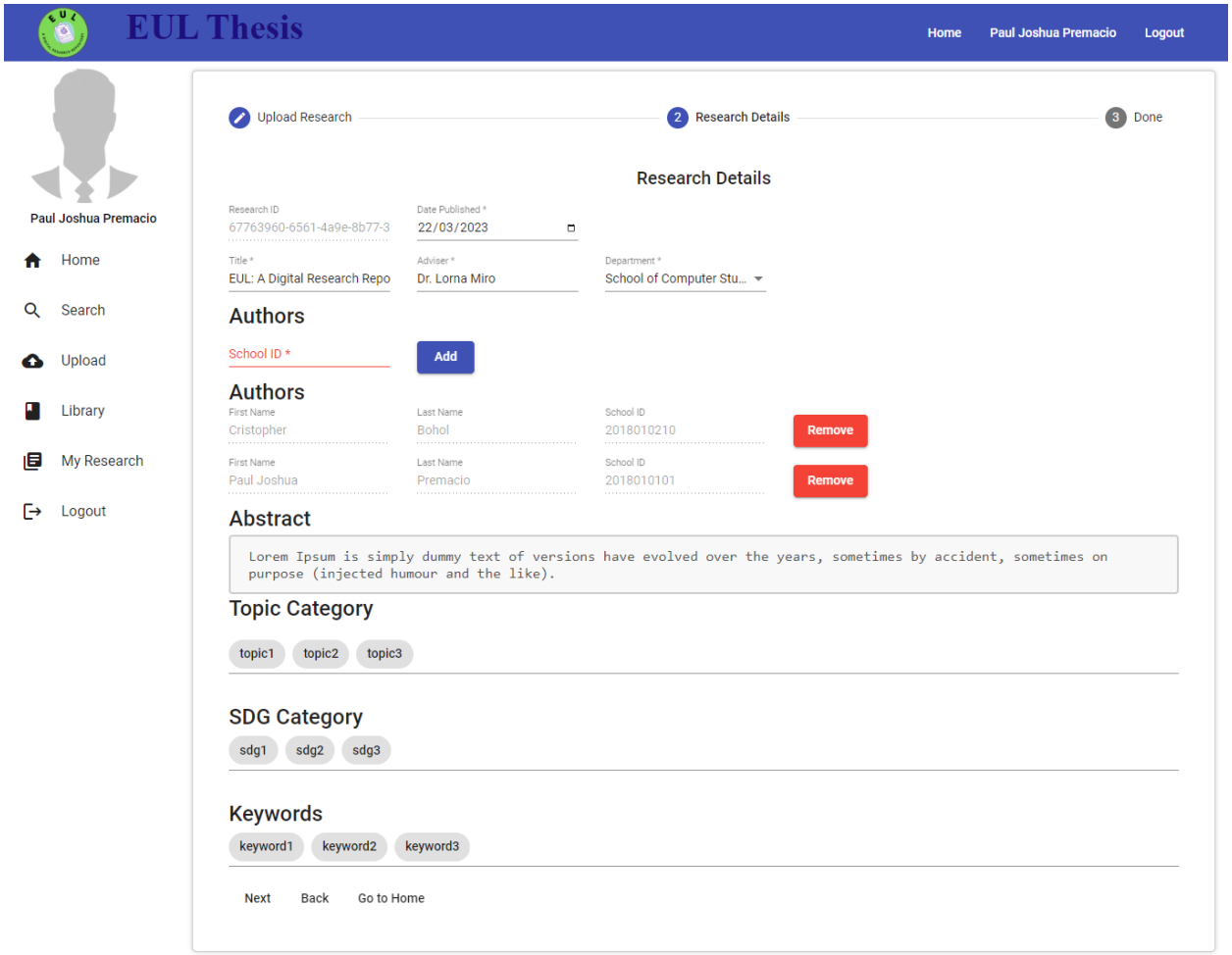
READ MORE

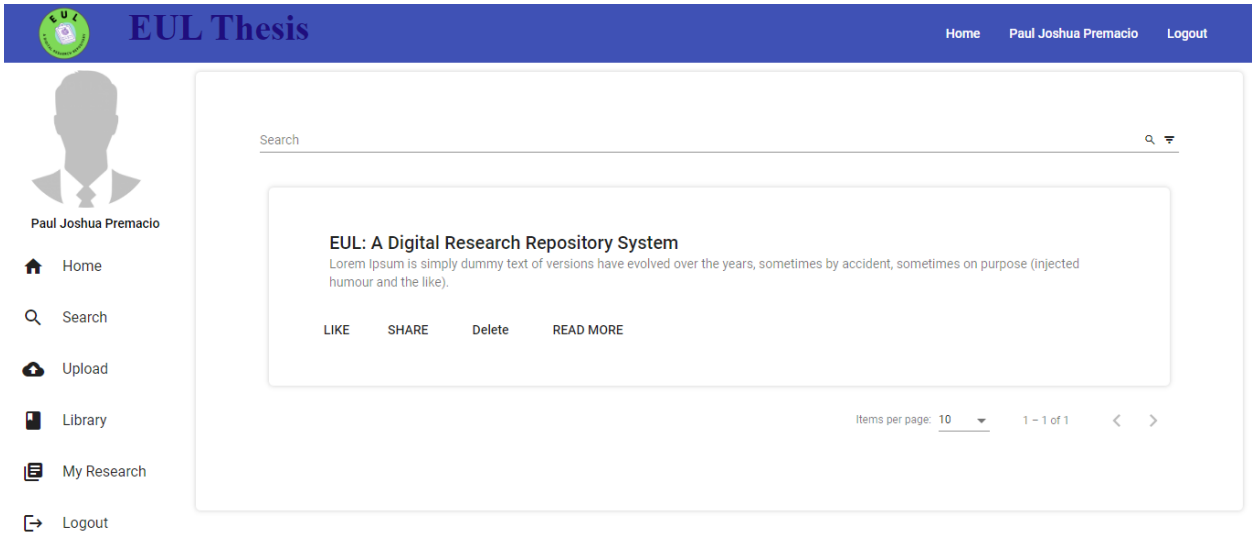
Vero iste consectetur aut atque.

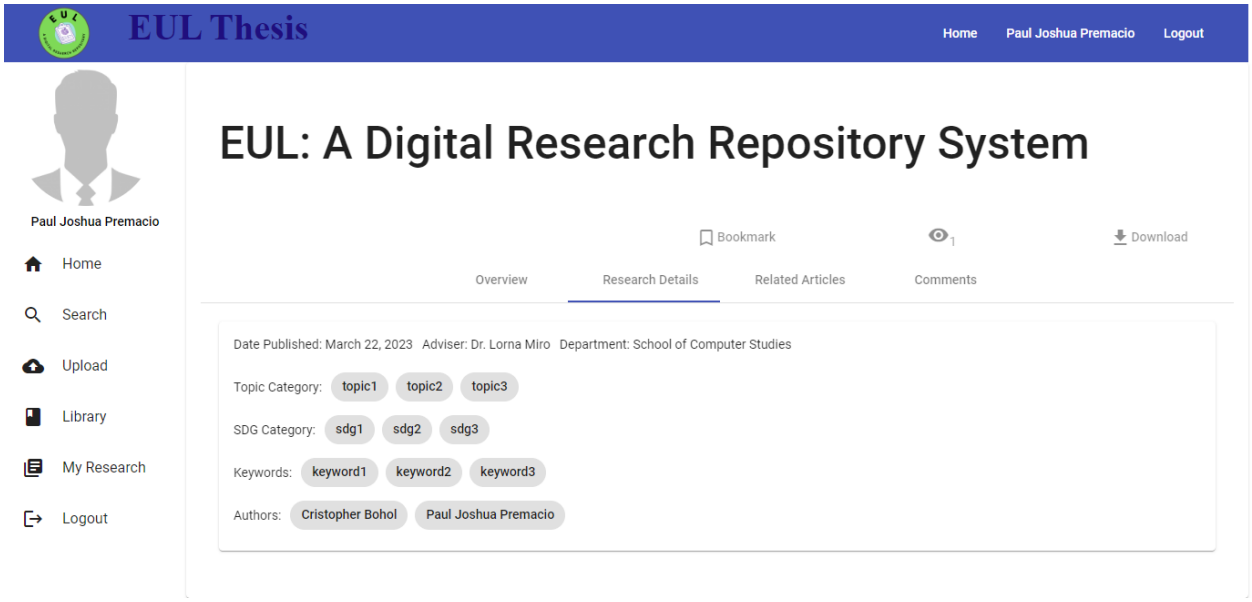
Nihil voluptatem amet. Aliquam rem vero. Eum odio necessitatibus. Dolores perspiciatis magnam. Perferendis repellendus corporis. Nesciunt rerum molestias. Nam sit labore. Suscipit neque consequuntur;

READ MORE

Items per page: 10 1 – 2 of 2


EUL Thesis
Home
Paul Joshua Premacio
Logout
Paul Joshua Premacio
Home
Search
Upload
Library
My Research
Logout
Upload Research
2 Research Details
3 Done
Research Details
Research ID
67763960-6561-4a9e-8b77-3
Date Published *
22/03/2023
Title *
EUL: A Digital Research Repo
Adviser *
Dr. Lorna Miro
Department *
School of Computer Stu...
Authors
School ID *
Add
Authors
First Name
Cristopher
Last Name
Bohol
School ID
2018010210
Remove
First Name
Paul Joshua
Last Name
Premacio
School ID
2018010101
Remove
Abstract
Lorem Ipsum is simply dummy text of versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).
Topic Category
topic1
topic2
topic3
SDG Category
sdg1
sdg2
sdg3
Keywords
keyword1
keyword2
keyword3
Next
Back
Go to Home


EUL Thesis
Home
Paul Joshua Premacio
Logout
Paul Joshua Premacio
Home
Search
Upload
Library
My Research
Logout
Search
EUL: A Digital Research Repository System
Lorem Ipsum is simply dummy text of versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).
LIKE
SHARE
Delete
READ MORE
Items per page: 10
1 – 1 of 1


EUL Thesis
Home
Paul Joshua Premacio
Logout
Paul Joshua Premacio
Home
Search
Upload
Library
My Research
Logout
EUL: A Digital Research Repository System
Bookmark
1
Download
Overview
Research Details
Related Articles
Comments
Date Published: March 22, 2023 Adviser: Dr. Lorna Miro Department: School of Computer Studies
Topic Category: topic1 topic2 topic3
SDG Category: sdg1 sdg2 sdg3
Keywords: keyword1 keyword2 keyword3
Authors: Cristopher Bohol Paul Joshua Premacio

CHAPTER III

SOFTWARE DEVELOPMENT AND TESTING

This chapter describes the implementation of the project in development, the various tools used to create and run the application, and testing methods to evaluate the question answering results. It contains sections for the Development and Testing Process.

DEVELOPMENT AND TESTING PROCESS

The system can be accessed via mobile and web platform. Each platform has its own technologies implemented. The process is a hybrid of Agile and Prototyping Model. An initial prototype is created and is refined as the development progressed.

For the web platform, these are the tools and development environment used:

- Visual Studio Code - is a light-weight IDE commonly used by developers. This is used to run the REST api services and the web frontend.
- Node.js - is a javascript framework used in this system as the backend environment.
- Angular 14 - a javascript framework used by this system as its frontend technology.

For the mobile platform, these are the tools and development environment used:

- Flutter Framework 3.3.7 - a relatively new framework for developing mobile, web, and desktop applications developed by Google. This framework is using Dart as its main language.
    - http v0.13.4 - an open source package by flutter used primarily in handling http requests.
    - json_serializable v6.3.1 - an open source package for easy handling of JSON, example converting JSON to Maps, or Maps to JSON.
    - path_provider v2.0.11 - an open source package for accessing and finding common accessed location in the file system.

- Android Studio version 2021.2 - is used by this system to run the emulator for debugging and testing.
- Visual Studio Code - the primary IDE used by the developer for fast and smooth development.

For the Python Backend, these are the tools and development environment used:

- Visual Studio Code - IDE used for developing and running python scripts.
- python 3.10.5 - the language used for creating the algorithms used by this system.
- Python Flask 2.2.3 - a python framework for building web applications. This system used the framework in serving the python scripts to be used by the web and mobile platforms in the form of http requests.
- nltk 3.7 - Natural Language Toolkit, a package used for handling all features related to natural language processing like lemmatization, POS Tagger, etc.
- pandas 1.4.3 - a python library for handling data analysis and structures. This library is used by the system primarily for converting dictionaries to CSV files.
- pdfplumber 0.7.5 - a python library for handling all tasks related to PDF files.

For Database Server,

- MySQL Workbench - digital filing cabinet for information, helps create, edit, view tables, run queries, manage user accounts, and server settings. Useful for database admins and developers.

Development Process

This part describes the technical process during the development. Figure 1 shows the general overview of the classification. There are different stages in the process of classification.

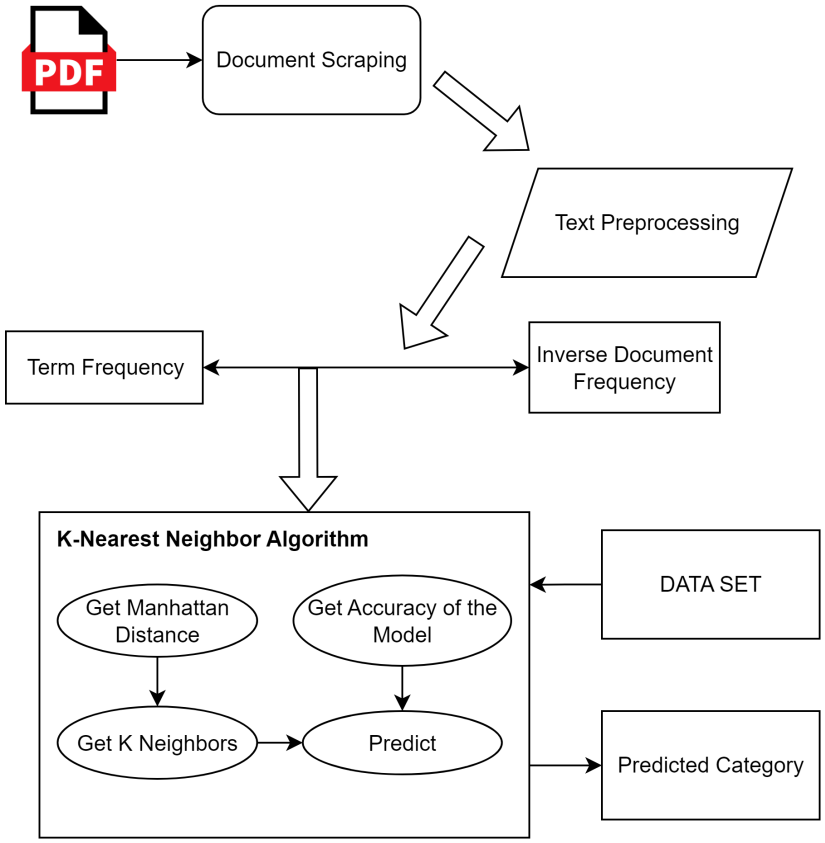


Figure 1 shows the overview of the classification algorithm.

Document Scraping

Commonly known as data scraping is a process of collecting data for analysis and prediction. In the system's case, this is used to collect all texts from the pdf file using a python pdf scraper.

Text Preprocessing

Data collected from scraping are in very large amount, and most of these are irrelevant and it adds too much noise for the data set. Preprocessing is one of the key steps before training the model. Figure 2 shows the process of cleaning the data.

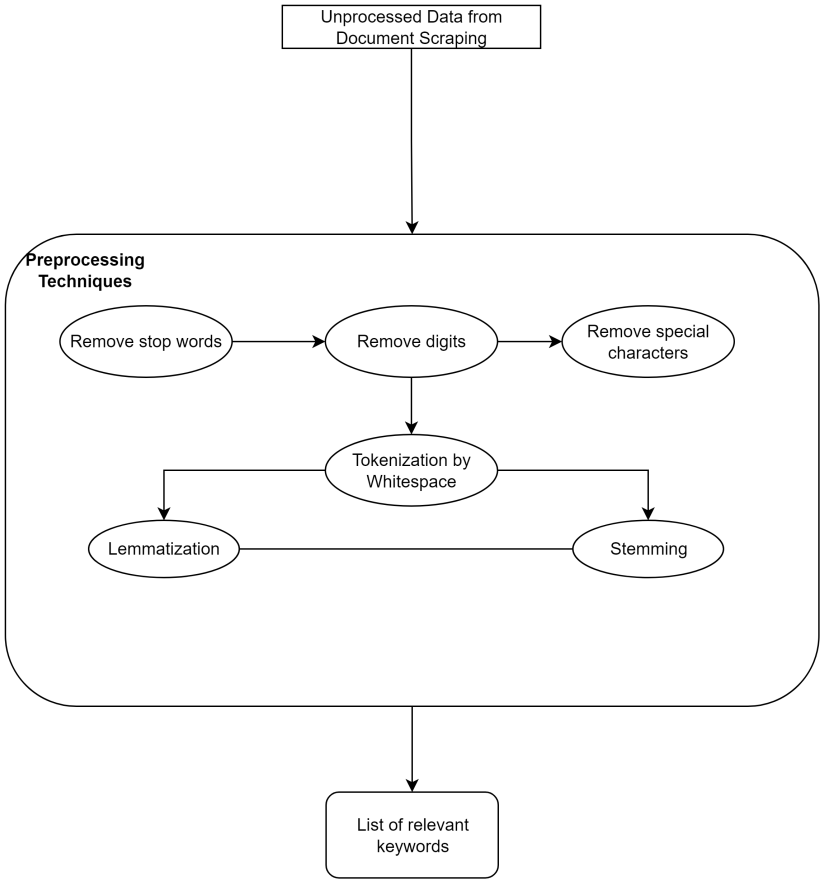


Figure 2 shows the overview of the preprocessing stage.

Term Frequency - Inverse Document Frequency

TF-IDF is used for text vectorization. In the system's perspective this is used as the data in feeding the model. Figure 3 shows the overview of the TF-IDF

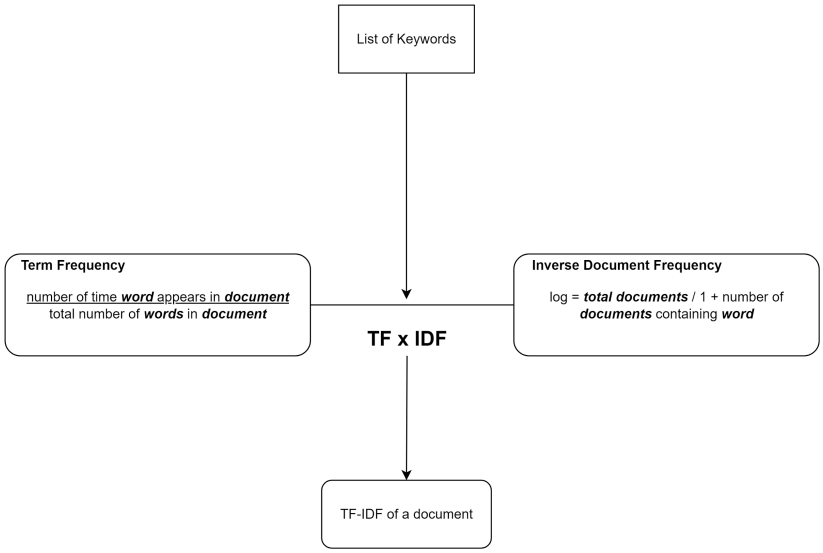


Figure 3 shows the TF-IDF model.

Document Classification

K-Nearest Neighbor Algorithm

Also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

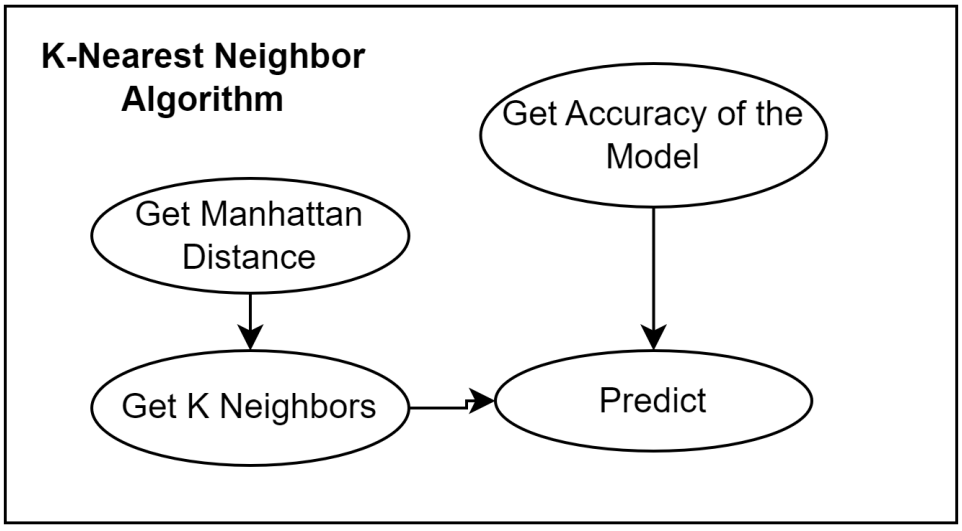


Figure 4 shows the process inside the model.

CHAPTER IV

SUMMARY, CONCLUSION, AND RECOMMENDATIONS SUMMARY OF FINDINGS

CONCLUSION

RECOMMENDATIONS

The K-nearest algorithm is not designed to analyze huge chunks of data and therefore is not fitted

//TODO

Some Disadvantages of KNN

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory – need to store all of the training data
- Given that it stores all of the training, it can be computationally expensive

// Deploy to net for easy access

// Disad

1. Include other stuff for classification, abstract, introduction, research method
2. Use other algorithms for classification other than KNN
3. Maybe subscription? IDK
4. Use other text vectorization, Word2Vec? Not TFIDF

REFERENCES

A. Online Website Resources

[1] History of artificial intelligence - javatpoint. www.javatpoint.com. (n.d.). Retrieved July 25, 2022, from https://www.javatpoint.com/history-of-artificial-intelligence

[2] Editor. (2021, November 17). *Document classification with machine learning: Computer vision, OCR, NLP, and other techniques*. AltexSoft. Retrieved August 4, 2022, from https://www.altexsoft.com/blog/document-classification/

[3] Burns, E. (2021, March 30). *What is machine learning and why is it important?* SearchEnterpriseAI. Retrieved July 25, 2022, from https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

[4] Petersson, D. (2021, March 26). *What is supervised learning?* SearchEnterpriseAI. Retrieved August 4, 2022, from https://www.techtarget.com/searchenterpriseai/definition/supervised-learning

[5] *Classification algorithm in Machine Learning - Javatpoint*. (n.d.). Retrieved August 4, 2022, from https://www.javatpoint.com/classification-algorithm-in-machine-learning

[6] Pratt, M. K. (2020, July 8). *What is unsupervised learning?* SearchEnterpriseAI. Retrieved July 28, 2022, from https://www.techtarget.com/searchenterpriseai/definition/unsupervised-learning

[7] Webb, G. I. (1970, January 1). *Naïve bayes*. SpringerLink. Retrieved July 27, 2022, from https://link.springer.com/10.1007%2F978-0-387-30164-8_576

[8] (LEDU), E. E. (2018, September 12). *Understanding K-means clustering in machine learning*. Medium. Retrieved August 4, 2022, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[9] Chen, S. (2020, May 26). *Getting started with text vectorization*. Medium. Retrieved August 1, 2022, from https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685#:~:text=Text%20Vectorization%20is%20the%20process,(L1)%20Normalized%20Term%20Frequency

[10] *Understanding TF-ID: A simple introduction*. MonkeyLearn Blog. (2019, May 10). Retrieved August 3, 2022, from https://monkeylearn.com/blog/what-is-tf-idf/#:~:text=TF%2DIDF%20

[11] By: IBM Cloud Education. (n.d.). *What is natural language processing?* IBM. Retrieved August 4, 2022, from https://www.ibm.com/cloud/learn/natural-language-processing

[12] *What is optical character recognition (OCR)?* IBM. (n.d.). Retrieved July 28, 2022, from https://www.ibm.com/cloud/blog/optical-character-recognition

[13] Kim, S.-W., & Gil, J.-M. (2019, August 26). *Research paper classification systems based on TF-IDF and LDA Schemes - human-centric computing and Information Sciences*. SpringerOpen. Retrieved August 4, 2022, from https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0192-7

[14] Mohsen Taheriyan University of Southern California, Taheriyan, M., California, U. of S., Nebraska, U. of, Army, U. S., Massachusetts, U. of, Oklahoma, U. of, Saic, Maryland, U. of, Nasa, & Metrics, O. M. V. A. (2011, August 1). *Subject classification of research papers*

based on Interrelationships Analysis: Proceedings of the 2011 Workshop on Knowledge Discovery, modeling and Simulation. ACM Conferences. Retrieved July 30, 2022, from https://dl.acm.org/doi/10.1145/2023568.2023579

[15] Maheshwari, A. (2018, July 17). *Report on text classification using CNN, RNN & Han*. Medium. Retrieved August 28, 2022, from https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f#:~:text=Text%20Classification%20Using%20Convolutional%20Neural,inspired%20by%20animal%20visual%20cortex.