


**BUZZMATCH: FILE CONTENT-BASED WITH IMAGE AND TEXT FILTERING  
CATEGORIZATION APPLICATION THROUGH KEYWORD MATCHING  
USING MODIFIED ARTIFICIAL BEE COLONY (ABC) APPROACH**

**A Thesis Presented to the  
Faculty of the College of Science  
Technological University of the Philippines  
Manila**

**In Partial Fulfillment of the  
Requirements for the Degree  
Bachelor of Science in Computer Science**

**By:  
ESPINOSA JOESEF ANDREI  
RIGA, RASHEED  
TORRES, JAN CHRISTIAN  
VALLOYAS, JON REXZEL**


**June 2024**

|   |   |              |                   |
|---|---|--------------|-------------------|
|  | <b>TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES</b><br>Ayala Blvd., Ermita, Manila, 1000, Philippines   Tel No. +632-5301-3001 local 608<br>Fax No. +632-8521-4063   Email: cos@tup.edu.ph   Website: www.tup.edu.ph | Index No.    | TUPM-F-COS-16-TAU |
|   |   | Revision No. | 00                |
|   |   | Date         | 07012022          |
|   |   | Page         | 1 / 1             |
| <b>THESIS APPROVAL SHEET FOR THE UNDERGRADUATE PROGRAMS OF THE COS</b>            |   |              |                   |


This thesis hereto entitled:


**BUZZMATCH: FILE CONTENT – BASED WITH IMAGE AND TEXT FILTERING  
CATEGORIZATION APPLICATION THROUGH KEYWORD MATCHING USING MODIFIED  
ARTIFICIAL BEE COLONY (ABC) APPROACH**


prepared and submitted by **JOSEF ANDREI ESPINOSA, RASHEED RIGA, JAN CHRISTIAN TORRES, and JON REXZEL VALLOYAS** in partial fulfillment of the requirements for the degree **BACHELOR OF SCIENCE IN COMPUTER SCIENCE** has been examined and is recommended for approval and acceptance.


  
**PROF. EDWARD CRUZ**  
 Adviser

Approved by the Committee on Oral Examination with a grade of **PASSED** on **JUNE 3, 2024**.

  
**PROF. ARIEL TOMAGAN**  
 Member

  
**PROF. MAY GARCIA**  
 Member

  
**PROF. JAN ELBERT L. LEE**  
 Member

  
**PROF. DOLORES L. MONTESINES**  
 Department Head/Chair

Accepted in partial fulfillment of the requirements for the degree **BACHELOR OF SCIENCE IN COMPUTER SCIENCE**.

Date: June 13, 2024

  
**DR. JOSHUA T. SORIANO**  
 Acting Dean

|                |                                   |
|----------------|-----------------------------------|
| Transaction ID | TUPM-COS-TAU-ELS-07012022 -0258PM |
| Signature      |                                   |

Transaction ID Legend: TUPX-AAA (Office Code)-BBB (Type of Transaction)-CCC (Initial of employee)-MMDDDDYYYY (month day year)-HHMMAM/PM (hourminutesAM/PM)

## ACKNOWLEDGEMENT

The researchers sincerely expressed their deepest gratitude to all who the individuals and institutions that have been instrumental in the completion of this research. First and foremost, the researchers would like to their heart to God almighty that gave us wisdom and strength throughout completing this study.

The researcher extends their heartfelt thanks to Professor Edward Cruz, who served as the thesis advisor for this study. His wisdom and guidance were valuable throughout the research process. He provided critical insights that shaped the direction of the study, and his suggestions for revisions following the thesis proposal were instrumental in completing the research. Without his wisdom and guidance, this study would not have been possible.

The researchers extend their thanks to Professor Jan Eilbert Lee, that made the researchers came up with the idea in this research, as well as his knowledge and recommendations to help the researchers complete the application.

The researchers also extend their sincere thanks to their family and friends for their unwavering support, motivation, and inspiration. Their encouragement and assistance in providing the tools and necessities required to complete the application were crucial. Additionally, the researchers deeply appreciate all the respondents who dedicated their time to evaluating the system and offering valuable suggestions for further improvement.

Overall, gratitude is expressed to everyone who has been involved in this throughout the completion of this study.

## **ABSTRACT**

This study focuses on providing assistance in efficiently categorizing and organizing PDF, MS Docx and TXT files based on image and text filtering. The researchers of this study found that there are professionals such as professors and students that have a hard time finding and retrieving electronic text-based documents. The main objective of this study is to develop a computer application that provides the user with categorizing electronic text-based documents by content. The system was developed using python and Tkinter framework, The system can automatically read and categorize the content of the documents. The evaluation criteria were based on ISO 25010 and the application is tested in terms of functional suitability, performance efficiency, usability, reliability, and maintainability. The level of acceptability of the system was graded with a mean rating of 3.53 which was interpreted as highly acceptable.

## TABLE OF CONTENTS

|   | <b>Page</b> |
|---|-------------|
| Title Page                                      | i           |
| Approval Sheet                                  | ii          |
| Acknowledgement                                 | iii         |
| Abstract  | iv          |
| Table of Contents                               | v           |
| List of Tables                                  | vii         |
| List of Figures                                 | viii        |
| List of Appendices                              | ix          |
| <br>  |             |
| <b>CHAPTER 1    THE PROBLEM AND ITS SETTING</b> | <b>1</b>    |
| Introduction                                    | 1           |
| Background of the Study                         | 2           |
| Objectives of the Study                         | 5           |
| Scope and Limitations of the Study              | 6           |
| Significance of the Study                       | 8           |
| <b>CHAPTER 2    CONCEPTUAL FRAMEWORK</b>        | <b>10</b>   |
| Review of Related Literature                    | 10          |
| Review of Related Studies                       | 30          |
| Conceptual Model of the Study                   | 32          |
| Operational Definition of Terms                 | 33          |
| <b>CHAPTER 3    METHODOLOGY</b>                 | <b>35</b>   |

|                  |  |           |
|------------------|--|-----------|
|                  | Project Design   | 35        |
|                  | System Design  | 35        |
|                  | Project Development  | 37        |
|                  | Operation and Testing Procedure                            | 39        |
|                  | Evaluation Procedure                                       | 41        |
| <b>CHAPTER 4</b> | <b>RESULT AND DISCUSSION</b>                               | <b>43</b> |
|                  | Project Description  | 43        |
|                  | Project Structure  | 43        |
|                  | Project Capabilities and Limitation                        | 47        |
|                  | Test Result  | 48        |
|                  | Evaluation Result  | 49        |
| <b>CHAPTER 5</b> | <b>SUMMARY OF FINDINGS, CONCLUSION AND RECOMMENDATIONS</b> | <b>55</b> |
|                  | Summary of Findings  | 55        |
|                  | Conclusion   | 55        |
|                  | Recommendation   | 56        |
|                  | <b>REFERENCES</b>  | <b>69</b> |
|                  | <b>RESEARCHER'S PROFILE</b>                                | <b>75</b> |

**LIST OF TABLES**

| <b>Table</b> |   | <b>Page</b> |
|--------------|---|-------------|
| 1            | Operating and Testing Procedure of the GUI Application (User Interface) | 40          |
| 2            | Likert Scale  | 42          |
| 3            | Test Modules Result   | 48          |
| 4            | Responses to Functional Suitability                                     | 49          |
| 5            | Responses to Performance Efficiency                                     | 50          |
| 6            | Responses to Usability  | 51          |
| 7            | Responses to Reliability  | 52          |
| 8            | Responses to Maintainability  | 52          |
| 9            | Overall Summary of Responses  | 53          |

## LIST OF FIGURES

| Figure |   | Page |
|--------|---|------|
| 1      | Behavior of honeybee foraging for nectar      | 22   |
| 2      | Pseudocode of Artificial Bee Colony Algorithm | 23   |
| 3      | Formula for Initialization Phase              | 24   |
| 4      | Formula Employed Bees Phase                   | 24   |
| 5      | Formula Fitness Value                         | 25   |
| 6      | Probability of Fitness Value                  | 25   |
| 7      | Conceptual Model of the Study                 | 32   |
| 8      | System Flowchart                              | 36   |
| 9      | Agile Software Development                    | 38   |
| 0      | Sample UI for the Application                 | 40   |
| 11     | Home Page Module                              | 44   |
| 12     | Text Categorizer Module                       | 45   |
| 13     | Instruction to use the Text Categorizer       | 46   |
| 14     | OCR Module                                    | 46   |



**LIST OF APPENDICES**

| <b>APPENDIX</b>                  | <b>Page</b> |
|----------------------------------|-------------|
| A      Survey Questionnaire Form | 57          |
| B      Grammarian Certificate    | 67          |
| C      Turnitin Report           | 68          |

## Chapter 1

### THE PROBLEM AND ITS SETTINGS

#### Introduction

Nowadays, E-text documents have become an important facet for sharing word processing and editing by mutual sharing and editing of documents. Although created and also viewed in electronic form, these electronic documents have several benefits over traditional paper-based documents. Electronic documents enhance accessibility of digital documents that can be accessed from nearly everywhere, better information retrieval through searching capabilities, and is giving more environmentally friendly because it minimizes the need for paper production.

Electronic, text-oriented document formats facilitate the processes of collaboration and sharing, as the document can be changed by more than one user and edited together—useful for people working together, in some cases in different parts of the world. Multimedia usage and interactive features add to the value of electronic text-based documents, making them a valuable tool for educational, professional, and personal purposes.

In summary, electronic text files have been discovering means of saving, explaining, and connecting information in the computer age. They are more accessible, easier to search, and better for the environment. Thanks to their collaborative and multimedia features, they have immense utility for businesses to schools, and as technology continues to progress, they will flourish as a resource used to consume information in the way people have come to enjoy. (Slettman, 2021).

## **Background of the Study**

The digital era made a major transformation in managing documents, from paper-based to electronic documents such as portable document formats or PDFs, and Word documents. These text-based electronic documents are now the standard for storing, sharing, and editing documents. This offers several advantages, including better accessibility, increased efficiency, and better file security (Rosano, 2023).

For centuries, paper documents have been the primary tool for storing and sharing information. However, due to the limitations of paper-based documents and the age-old problems associated with it: physical storage requirements, vulnerability to physical damage, and difficulties in manipulation, distribution, and retrieval, electronic document formats have been adopted. The Portable Document Format (PDF), developed by Adobe Systems in the early 1990s, is a file format used to present documents, in a manner independent of application software, hardware, and operating systems; a PDF has the capability to encapsulate text, fonts, images and vector elements of a document in a single file suitable for printing and screen viewing (Adobe Systems, n.d.). Microsoft Word serves as a versatile digital platform for document creation and editing, closely resembling the experience of working with paper documents. Users can enter and format text, mimicking the act of typing or writing on paper. It provides tools to control page layout and document structure, allowing for the organization of content into sections and headings. Similar to adding images or graphics to paper documents, Word enables the insertion of multimedia elements. It also simplifies the process of creating tables, charts, and other visual representations. After document creation, it facilitates printing,

offering various print settings like page selection and duplex printing. Microsoft Word also allows for file storage, archiving, and easy digital access. While it differs from paper in terms of collaboration and security features, it remains a powerful tool for creating and editing text-based documents that replicate the familiarity of working with traditional paper (Microsoft, n.d.).

Changing from paper documents to electronic text-based documents has the advantage of making them more accessible. This contrasts with paper which demands being at the specific location where the document resides, whereas electronic text-based documents can be distributed digitally through emails, in cloud storage or otherwise and still be used for collaboration and exchanging information across different points on earth (Martin & Shaw, 2021). Moreover, electronic text-based documents are supported by numerous devices like computers, tablets and smartphones that make access to information possible while one is in transit.

Another benefit of using electronic text-based documents is efficiency in document management. Traditional paper-based filing, organizing and retrieval processes take much time and are prone to human errors. However, electronic text-based documents have features like searchable texts, bookmarks and hyperlinks that enable quick and accurate navigation within these documents (Nguyen, 2023). Moreover, electronic systems allow encryption of different types of information to protect the integrity of the document through digital signatures. This kind of protection is very essential for a number of industries (Buch, 2020).

Apart from this, switching from conventional paper documents to electronic text ones promotes environmental sustainability. The production and disposal of paper documents have major eco-impacts including deforestation, energy use as well as waste generation. Thus, by adopting the use of electric text-based papers organizations or individual persons can be able to

reduce their papers consumption hence conserving natural resources as well as contributing towards mitigating climate changes (Latsoomanan, 2023).

The use of electronic text files in organizations and industries has increased steadily, which poses a problem about how to categorize these files according to their content. Since most people nowadays save documents in form of e-text on their computers, there has been the necessity for individual or organizational efforts to keep track of large numbers of files. Nevertheless, traditional techniques such as using file names or metadata are not practical because electronic text-based documents are visually consistent representations of paper documents. This is a challenge for users who need to locate specific electronic text-based documents quickly and accurately, leading to low productivity and potential data loss (Nielsen & Kaley, 2020).

This is due to the fact that electronic text-based document files are naturally images of real documents hence traditional classification methods that depend on file names or metadata become inadequate. Hence, it becomes laborious and time-consuming to derive useful information from electronic text-based document files. Consequently, users face challenges in organizing and searching for text-based electronic documents; this hampers their ability to make good use of accumulated data, thus preventing better workflow arrangement (MESHDS, 2021). Additionally, the problem is worsened by the fact that there are so many PDF documents being created and shared daily making it difficult to even categorize PDF files alone based on content. Recent statistics indicate an increasing popularity of PDF. In 2020, global PDF creation reached a massive milestone of 2.5 trillion, representing a substantial increase from previous years (Rajeev, 2021). It is clear that this exponential growth rate in using PDF creates a major issue for efficiently managing and organizing these records based on what they contain.

Further studies have shown that professionals and knowledge workers spend quite good amounts of time looking for information within their document repositories. A report by IDC (2016) indicates that approximately, employees spend 9.5 hours each week just searching for documents with an average success rate of retrieval being only 50%. Not only does this show the prevalence of the problem but also its impact on productivity and overall workplace efficiency.

The burden associated with categorizing and retrieving this immense number of electronic text-based documents shows that there is a need for effective and automated ways to categorize information. In this regard, the use of artificial intelligence (AI) in addition to machine learning algorithms can be the perfect solution towards lean document management processes that will enhance users' productivity levels.

Since it is difficult to classify electronic text files according to their contents, researchers have come up with a solution that uses AI for automation. Organizations can overcome challenges involved in manual categorization by developing artificial intelligence (AI) systems specifically designed to classify electronic text-based documents. Some methods used may include techniques like extracting information from electronic text-based documents, understanding what they entail or assigning suitable categories or tags among others. In addition, researchers will apply Swarm Intelligence to boost computer performance by running multiple agents at once. This automation saves time and energy while improving the precision and consistency in classification leading to better document handling and retrieval processes.

## **Objectives of the Study**

### ***General Objective***

The general objective of the study is to create Artificial Intelligence using Swarm Intelligence to organize and categorize the contents of electronic text-based documents.

### ***Specific Objectives***

The study has the following specific objectives:

1. Design a system with the following features:
  - a. Graphical User Interface (GUI):
    - i. Create an interactive UI for the users.
    - ii. Locate file destination.
    - iii. Ask for keywords.
  - b. Datasets
    - i. List of electronic text-based documents with random topics and contents
2. Create the system with developmental tools listed below:
  - a. Front-end Tools
    - i. Python
    - ii. Tkinter
  - b. Back-end Tools
    - i. Python
3. Test and evaluate the system using functional suitability and reliability.
4. Determine the acceptability level of the system using the ISO 25010 criteria in functional suitability, performance efficiency, usability, reliability, and maintainability.

### **Scope and Limitations of the Study**

The File Content-Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach is designed specifically for automated categorization of electronic text-based documents. The system focuses solely on utilizing the Artificial Bee Colony (ABC) as the core technique for the categorization process. Additionally, this research will assist in simulating the conduct of bees while making use of parallelization of processes with a multi-threading approach derived from foraging in bees.

The system utilizes Keyword Matching Approach where keywords extracted from electronic text-based documents form the basis for categorization. The ABC algorithm has been used to optimize the categorization process, which is inspired by Bee Foraging behavior. The algorithm efficiently and accurately classifies electronic text-based documents by iteratively exploring and exploiting the solution space.

Through the use of ABC algorithm, the system is able to dynamically adjust itself according to the features and difficulties of e-texts. It is through this process that the most suitable classes for each document are found based on keyword extraction. This method allows for efficient categorization while minimizing human intervention.

The system concentrates mostly on ABC algorithm which helps in specialized solution towards electronic text-based documents categorization. By using its optimization capabilities together with Keyword Matching approach, a system wishes to give effective and accurate outcomes in electronic text-based documents categorization as possible.

Overall, the File Content- Based with Image and Text Filtering Categorization Application Via Keyword Matching using Modified Artificial Bee Colony (ABC) Approach



delivered a specialized framework in which the ABC algorithm is applied utilizing a more efficiently to automate the categorization of the electronic copious number of text-based documents which can be a perfect solution in organizing documents and text retrieval.

File formats will be limited to the following: Standard file formats (txt) Microsoft office file formats (.docx), and PDFs.

The assessment instrument that will be used to evaluate the acceptability of the application consists of the ISO 25010 Software Quality Model.

### **Significance of the Study**

#### *For the Professors and Teachers*

The lack of teachers and professors are one of the problems in the Philippines, thus the schools and universities have no choice but to give teachers additional courses to teach. This can lead to a mixture of resources in their laptop. Using automated text-based categorization can help professors and teachers save time finding resources on their subjects.

#### *For the Students*

To save money, several students tend to find the pdf format of the books used by professors and teachers. Furthermore, having a lot of subjects can lead the students to lose their documents, thus, organizing the documents automatically can be very beneficial to the students.

#### *For the Researchers*

This research is important for researchers, as it will provide valuable insights and improve their proficiency in software development. The outcomes of this study are expected to

contribute to the body of knowledge, enhancing the researchers' understanding. Furthermore, the detailed investigation is anticipated to refine their skills in software development, thereby aligning them with current advancements in the field.

## **Chapter 2**

### **CONCEPTUAL FRAMEWORK**

This chapter provides an overview of related literature, related studies, the conceptual model of the study, and the operational definition of terms relevant to this study.

#### **Review of Related Literature**

This section presents key concepts and ideas on the topic of the study.

#### **Text – Based Documents**

Text-based document is any document that consists of text characters primarily, rather than documents that tend to have a significant amount of non-text elements which include images, graphics and multimedia content. Text-based documents can be created and edited with different types of text editors or word processing software (Adobe Systems, n.d.).

#### **Types of Text – Based Documents**

A file format that is widely used, the portable document format (PDF) is designed for consistent presentation of documents across various platforms. PDFs are mainly text-based since they can preserve textual content well in spite of allowing non-text elements to be incorporated therein. This format contains text, fonts, images and media objects in one file so as to enhance uniformity of documents regardless of devices or software used. PDFs enable easy selection, copying and searching of texts hence applicable in various areas including official reports, business letters, education materials and e-publishing. The significance of PDF's versatility in

today's digital sharing and dissemination contexts cannot be overemphasized especially within educational and professional settings (Cakir, 2016).

Within the sphere of current document production, a *Word document* is identified as a file that was created and edited with Microsoft Word, which is a modern word processing program from Microsoft Office suite. These documents have different things in them such as carefully formatted text, images, tables and graphs indicated by the extensions .doc and .docx. It is known for its powerful tools for text formatting, page layout adjustment and easy embedding of multimedia files among other functions (Microsoft Corporation). The user-friendly nature of Microsoft Word makes it ideal for creating different types of documents ranging from reports to resumes and letters among many others both professional and non-professional ones. Word documents are formal, feature-rich tools necessary for academic and professional work in contemporary research and documentation settings (Britannica, 2023).

A *text file* has no formatting or embedded styling such as bold or color fonts; it contains plain text characters alone without any other formatting styles. Simplicity and universal compatibility are its main advantages over other formats. By being easily readable by humans, this format poses several benefits in context to scholarly research especially when diverse texts processors are used to edit it (Rouse, 2016).

## **File Categorization**

File classification, file organization or categorization known file, is the systematic process of placing files into separate groups or categories on the basis of their attributes, contents and metadata. The aim of this process is to ensure that files can easily be managed, retrieved and organized within a digital system or storage environment (Gelavska, 2023).

Categorizing files aids in maintaining uniformity in naming conventions and file structures thereby streamlining collaboration and file sharing. It helps track and monitor files thus minimizing loss or misplacement of important documents. Files can be organized based on type, project, date or topic hence facilitating adaptable organization depending on the specific necessity. In short, it adds value to managing files that tend to make a business run smoothly (Kim, 2022).

### ***File Categorization Technique***

Numerous ways can be used to categorize files depending on an organizations or individual's specific needs and the context. The most common include chronological, alphabetical, assigning metadata tags or attributes, or implementing classification schemes based on predefined categories (Government of Tamil Nadu 2019). By employing these approaches, you will be able to keep your files in order and have a framework that is logical for managing digital assets.

The Texas State Library and Archives Commission highlights alphabetical-based categorization as one of the key techniques employed in the alphabetical filing system. In either hard copy or electronic form, this technique involves arranging files according to alphabet letters so as to make information easily accessible. By using file folders labeled A through Z and using this approach people are able to put together a well-organized system. Alphabetical filing is a viable option for document management with potential application across a wide range of uses from personal record keeping to professional environments (Sheahan n.d.).

The Entrepreneur Network has come up with the technique of chronological-based categorization to be the best method for business people to file and access records conveniently.

This is particularly useful when arranging files emanating from meetings. For instance, with the physical or electronic record, one can have folders dividing them into weeks, months, quarters or years depending on personal requirements which make it easy to keep and retrieve information. In summary an entrepreneur can improve his/her business operation by adopting a system of filing documents in a certain order of time as this would enhance proper records keeping as well as making it easier to find important papers when needed (Sheahan. n.d.).

Subject-based categorization helps in organizing files by their content, and it is beneficial to information search. Files with related themes are grouped together through keyword tagging hence making it easier for them to be located when needed. This flexible system does not require an exact chronology. Therefore, a more adaptable organization is possible. Several tags can be assigned to a single file as a way of linking the document to different categories or topics especially in the case of huge and diverse time-framed collection of files. Through tag searches, this method simplifies the process of finding and retrieving information thus saving time that could have been used otherwise for other activities aimed at boosting productivity levels. Moreover, it promotes fruitful teamwork and sharing ideas because any member of the group can access documents concerning specific subjects (Ashlstrom, 2005).

Organization purposes necessitate that the researchers employ subject-based categorization. In many devices, subject-based categorization is more pragmatic than alphabetical based and chronological based categorizations. To allow users to locate content on a specific topic or concept of their interest, they can be grouped together based on this. By doing so, the accessibility of the system is enhanced and its usability increased since information can be retrieved in an intuitive way and faster when it is most needed. Furthermore, subject based

categorization allows easier updating and expansion as new subjects come up because it accommodates the dynamic nature of knowledge.

### **Keyword Matching**

It's important to understand what keywords are all about before assessing their significance. The term 'keywords' also known as 'key phrases' in this connection is defined by Çano & Bojar (2019) as "a concise collection of one or a few words that encapsulate a concept or topic addressed in a document". For instance, these keywords act like street signs that point out to readers what they should pay attention to when reading materials such as essays or articles online.

Tomokiyo and Hurst (2003) believed that for keywords to be effective they must be "informative". Informative in this context refers to the chosen keywords having some informational components associated with background knowledge or domain expertise. By selecting informative keywords, one ensures that the subject matter is more accurately represented by capturing main ideas as well as essential details of a document thereby aiding its overall comprehension and easy retrieval.

Similarly, Bharti et al. (2017) argue that "core sentiment" is another attribute of keywords according to Zhang (2008). In other words, they always portray the text's underlying emotions, tone or any perspective shown by it. This feature of sentiment becomes crucial in carrying out sentiment analysis tasks or when looking to arrange and classify materials based on their emotional content. Likewise, not only are keywords necessary for grasping as well as retrieving documents but SEO and digital marketing depend highly on them. In this case, having the right key words is important when seeking targeted traffic to your website besides improving

its prominence over search results. To achieve higher rankings from Google and other search engines therefore, a careful integration of these words into web page contents, metatags as well as other promotional stuff is thus vital for businessmen. Thus making it possible for a firm to target the appropriate audience hence efficiently reaching potential customers while increasing overall involvement

Texts can be organized and classified using keywords. Thus, investigators and data analysts cluster texts with the same themes or topics under such headings to enable easy retrieval and examination of much information involving this kind of themes or topics through text classification. Researchers also can find out key themes within their corpus by analyzing critically the key words in different texts; thus, guiding them to identify patterns or trends in complex data sets which would otherwise go unnoticed. In a word matching information retrieval system is very important as it involves identification and matching of specific keywords or phrases. For instance, search engines can compare user queries' keywords with indexed documents so as to efficiently pull relevant search results while ignoring irrelevant ones. Their research reveals that keyword match has greatly improved search engine performance and user satisfaction (Manning et al., 2008).

Keyword matching also assists in content analysis and classification tasks. For example, recognizing and linking specific words or phrases from a text enables researchers to easily arrange and interpret vast quantities of data. In Choudhury et al.'s (2014) study, sentiment analysis and opinion mining were found to be major areas where useful knowledge can be gotten through keyword matching of user-generated contents.



In summary, keywords are brief descriptions of the concepts or topics discussed in a document. They are expected to provide information that briefly encompass all the main points related to the background knowledge but reflecting those attitudes expressed in this text essence. Smart use of keywords is important for efficient search retrieval, good document organization and increased visibility across different domains. Besides, search engine performance improvement can happen if there is effective content analysis and classification through keyword matching which informs while extracting useful information from text data.

## **Swarm Intelligence**

Swarm behavior arises from self-organized and decentralized systems in which the collective intelligence of simple constituents emerges. These insects dwell in hives, mounds or colonies and can be seen working together on huge tasks without any central control but rather depending upon their individual interactions as well as simple rules. Swarm Intelligence holds that the total intelligence of a group is greater than the sum of its parts, thus highlighting collaboration and self-organization as they relate to intelligent global phenomena.

The ability to solve complicated problems through the interactions of individual agents is one of the main variables that make swarm intelligence quite an interesting field as it is inspired by collective behavior observed in nature. Here's an instance, birds flock and fish swim together in groups showing remarkable instincts and skill in coordinated movement and direction even when there is no command or direct communication taking place. According to Kennedy (2006), swarm intelligence is “collective intelligence emerging from decentralized decisions made by many ‘simple’ agents” that supersedes what an agent can do constitutively. Through

decentralizing decision-making, local interaction and self-organization emerge properties which are not displayed by one agent alone and tasks that go beyond any single agent's ability are attained. Robotics, optimization and artificial intelligence have been provided with new problem-solving methods in this kind of knowledgebase that provides different perspectives. .

Swarm intelligence is used in numerous arenas among them being optimization algorithms, robotics and distributed or even clustern computing. Here's a good example, swarm-intelligence-based algorithms like Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) are used to address optimization problems through simulating the behavior of natural systems. This is a swarm intelligence technique which optimizes solutions in intricate problem spaces by considering self-organization and collective behavior (Blum & Li, 2008).

### ***Types of Swarm Intelligence***

Particle Swarm Optimisation (PSO), for example, mimics how schools of fish behave or how birds flock together. This was first cited by Kennedy and Eberhart (1995). It works by sending single entities called particles flying around different points trying to locate optimal solutions

Each particle in the swarm represents a potential solution to the optimization problem and so can move around in search space by changing its position and velocity. This means that they are constantly altering their positions relative to each other until eventually locating optimum values. The particle's position is updated based on its own personal best-known position as well as with respect to the existing global optimum common to all other particles.

The PSO technique is used widely across many disciplines such as engineering designs, data mining activities, picture manipulation and neural network training. Single-objective problems have been solved using this method, but also multi-objective optimization cases were successfully addressed through it.

To resemble ants as they search for food, **Ant Colony Optimization (ACO)** is a Meta heuristic algorithm. This method is widely used to solve optimization problems especially in combinatorial optimization. Imitating ant trails of pheromones which are used by ants to communicate and navigate within their environment in the colony, ACO emulates this phenomenon. Pheromone trails also known as trail markers on the edges of the problem graph that artificial ants travel through to reach other and hence collectively arrive at good solutions.

It was Marco Dorigo who first suggested ACO in 1992 and it has attracted much attention since then resulting into numerous applications to various real-world problems. Marco's initial study on ACO appeared in a paper he authored entitled "Optimization, Learning and Natural Algorithms" (Dorigo & Stutzle, 2006) The technique has been very successful in solving hard combinatorial optimization problems such as traveling salesman problem (TSP), vehicle routing problem (VRP), job scheduling problem etc.

ACO is an approach that has been created to design pheromone trails, so as to make positive feedback loops. The ants prefer the shortest paths because they have more pheromones deposited on them. As a result, many ants go through a solution space, leave more pheromones behind at better solutions and by doing this do not allow others to discover it.

ACO is widely used in various sectors such as transport, telecommunication, logistics or scheduling. For example, within the telecommunications field ACO was used for network

routing optimization, channel allocation and resource allocation problems (Gravett 2017). In logistics ACO has been employed for vehicle routing problems (Amodeo et al., 2017) and scheduling in transportation networks owned by different firms which are competing for clients. It is valuable in situations where classical algorithms are not effective enough to find optimal solutions because of its effectiveness and versatility in solving optimization problems.

***Fish Swarm Optimization*** (FSO) algorithm is an optimization technique based on the population which is motivated by schools of fish. This algorithm models interconnections between fish individuals and their surroundings so as they can search for optimal solutions together. Every single fish represents a potential solution for this optimization problem, and it continually updates their positional coordinates and movements via some rules (Azizi 2014).

The movement of fishes in FSO algorithm depends on different things like the position of the current best solution (“food”), the positions of neighboring fishes as well as distances to boundaries of searching spaces. The algorithm includes several strategies such as collective behavior, prey-predator behavior, random exploration so that it may balance exploration with exploitation during optimization (Mu et al 2016).

FSO has been successfully applied to a wide range of optimization problems, including function optimization, image segmentation, and clustering. The algorithm has shown great results compared to other metaheuristic algorithms in terms of speed and solution quality.

### **Artificial Bee Colony**

ABC is a kind of metaheuristic optimization algorithm which borrows some ideas from honeybee’s food searching behavior. It is an effective method of solving difficult

optimization problems. In this process, ABC simulates the behavior of bees in the hive such as employed bees, onlooker bees, and scout bees to find an optimal solution within a search space (Karaboga, 2005).

ABC algorithm has become one of the most famous optimization algorithms due to its ease in applying in different systems and also suitable for different forms of optimization problems. Notable applications include engineering design, data mining for large machine learning models and image processing among others.

### ***Foraging Behavior of Honeybees***

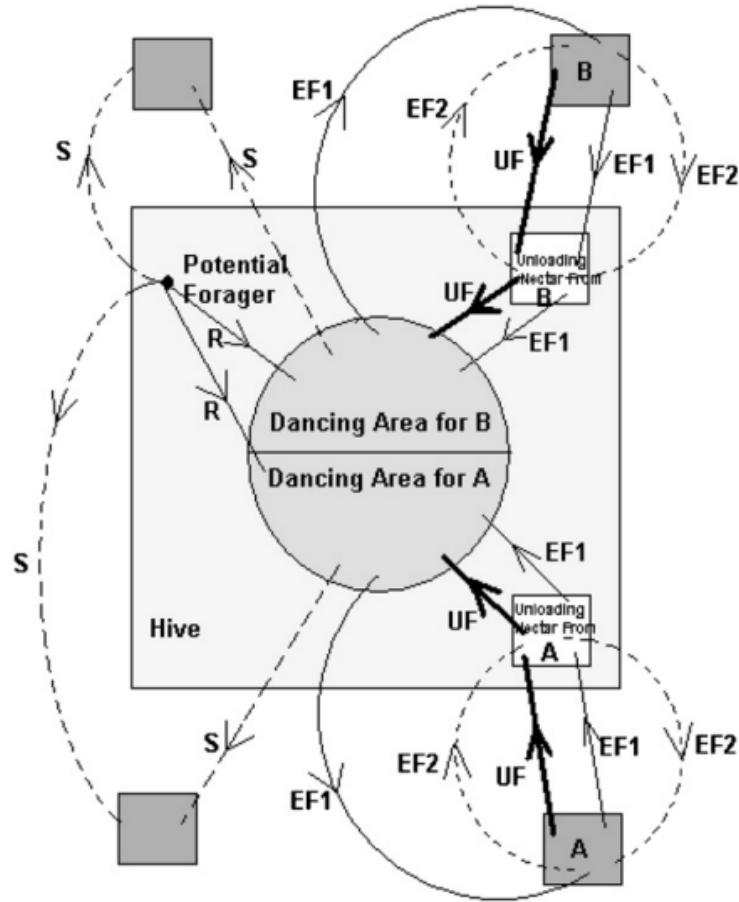
The fundamental model for understanding the emergence of collective intelligence in honeybee swarms focuses on three key components: food sources, employed foragers, and unemployed foragers. This minimal model outlines two primary behavioral modes: recruitment to a nectar source and abandonment of a source. These components and modes form the basis of studying how honeybees collectively make forage selection decisions and exhibit intelligent behaviors as a swarm (Jeanne, 1986).

***Food sources*** are evaluated based on various factors, including their proximity to the nest, the concentration of energy they provide, and the ease of extracting that energy. To simplify the evaluation process, the value or "profitability" of a food source can be represented using a single quantity

***Employed foragers*** are those who are connected to a specific food supply that they are currently utilizing. They carry information about this specific source, its proximity to the nest, its orientation, and its profitability, and they share this information with a certain probability.

*Unemployed foragers* are looking for a food supply to take advantage of. Unemployed foragers can be divided into two groups: scouts who explore for new food sources in the area around the nest and onlookers who wait in the nest and locate a food source using the knowledge supplied by employed foragers (Seeley, 1995).

The exchange of information among bees plays a vital role in the formation of collective knowledge within a hive. When observing a hive, certain elements can be consistently identified across all hives. Among these, the dancing area holds utmost significance as it facilitates the exchange of information. Bees communicate about the quality of food sources through a specific type of dance known as the waggle dance. On the dance floor, an onlooker bee has access to information about all the currently rich food sources. By observing multiple dances, the onlooker can select the most profitable food source to exploit. The availability of more information about highly profitable sources increases the likelihood of onlookers choosing those sources. Employed foragers share information about food sources based on their profitability, and the duration of waggle dances conveying this information is longer. As a result, the recruitment of foragers to a particular food source is proportional to its profitability (Loengarov & Tereshko, 2005).



**Figure 1.** Behavior of honeybee foraging for nectar.

Within the context of the artificial bee colony model, there are two possible roles for a bee. Firstly, it can act as a scout, independently venturing out from the hive to search for food based on internal motivations or external cues. This bee initiates exploration spontaneously ('S' in Fig. 1). Secondly, a bee can become a recruit after observing the waggle dances performed by other foragers. Inspired by the dances, the recruit bee sets off to search for the indicated food source ('R' in Fig. 1). Once the food source is found, the bee uses its memory to exploit the location and becomes an "employed forager." Upon returning to the hive, the employed forager unloads the collected nectar into a food store. At this point, the bee has several options: it may abandon the food source and become an uncommitted follower (UF), it may recruit other nest

mates by performing a dance and then return to the same food source (EF1), or it may continue foraging at the food source without recruiting other bees (EF2).

It is worth noting that not all bees commence foraging simultaneously. Observations and experiments have revealed that new bees begin foraging at a rate proportional to the difference between the eventual total number of bees and the number of bees currently engaged in foraging activities.

### ***Artificial Bee Colony Algorithm***

In ABC, employed bees explore the search space by exploiting the information from the best solutions discovered so far. Onlooker bees select promising solutions based on the quality of food sources. Scout bees are responsible for introducing random exploration by searching for new solutions. Through the iterative process of employed bee phase, onlooker bee phase, and scout bee phase, the ABC algorithm aims to converge towards the global optimum.

```
Initialization Phase
REPEAT
    Employed Bees Phase
    Onlooker Bees Phase
    Scout Bees Phase
    Memorize the best solution achieved so far
UNTIL(Cycle=Maximum Cycle Number or a Maximum CPU time)
```

**Figure 2.** Pseudocode of Artificial Bee Colony Algorithm



**Initialization Phase:** each solution vector in the population of food sources (denoted as  $xm \rightarrow$ , where  $m$  ranges from 1 to  $SN$ ,  $SN$  being the population size) is initialized by scout bees. Control parameters are also configured during this phase. Since each food source,  $xm \rightarrow$ , is a solution vector to the optimization problem, each  $xm \rightarrow$  vector holds  $n$  variables,  $(xmi, i=1 \dots n)$ , which are to be optimized to minimize the objective function.

$$x_{mi} = l_i + rand(0, 1) * (u_i - l_i)$$

**Figure 3.** Formula for Initialization Phase

where  $l_i$  and  $u_i$  are the lower and upper bound of the parameter  $xmi$ , respectively.

**Employed Bees Phase:** Employed bees search for new food sources ( $vm \rightarrow$ ) having more nectar within the neighborhood of the food source ( $xm \rightarrow$ ) in their memory. They find a neighboring food source and then evaluate its profitability (fitness). For example, they can determine a neighbor food source  $vm \rightarrow$  using the formula given in Figure 4.

$$v_{mi} = x_{mi} + \phi_{mi}(x_{mi} - x_{ki})$$

**Figure 4.** Formula Employed Bees Phase

where  $xk \rightarrow$  is a randomly selected food source,  $i$  is a randomly chosen parameter index and  $\phi_{mi}$  is a random number within the range  $[-a, a]$ . After producing the new food source  $vm \rightarrow$ , its fitness is calculated, and a greedy selection is applied between  $vm \rightarrow$  and  $xm \rightarrow$ .

The fitness value of the solution,  $fitm(xm \rightarrow)$ , might be calculated for minimization problems using the following figure 5.

$$fit_m(\vec{x}_m) = \begin{cases} \frac{1}{1 + f_m(\vec{x}_m)} & \text{if } f_m(\vec{x}_m) \geq 0 \\ 1 + abs(f_m(\vec{x}_m)) & \text{if } f_m(\vec{x}_m) < 0 \end{cases}$$

**Figure 5.** Formula Fitness Value

where  $f_m(xm \rightarrow)$  is the objective function value of solution  $xm \rightarrow$ .

**Onlooker Bees Phase:** The unemployed bees in the Artificial Bee Colony (ABC) algorithm can be further divided into two groups: onlooker bees and scouts. Within the algorithm, employed bees communicate information about food sources to the onlooker bees. The onlooker bees, waiting in the hive, make their food source choices based on the information received from employed bees. The selection of a food source by an onlooker bee is determined probabilistically, utilizing probability values calculated from fitness values provided by employed bees. In this process, a common technique used for selecting food sources based on fitness is the roulette wheel selection method (Goldberg, 1989).

The probability value  $p_m$  with which  $xm \rightarrow$  is chosen by an onlooker bee can be calculated by using the expression given in Figure 6.

$$p_m = \frac{fit_m(\vec{x}_m)}{\sum_{m=1}^{SN} fit_m(\vec{x}_m)}$$

**Figure 6.** Probability of Fitness Value

After a food source  $xm \rightarrow$  for an onlooker bee is probabilistically chosen, a neighborhood source  $vm \rightarrow$  is determined by using figure 4. and its fitness value is computed. As in the employed bees phase, a greedy selection is applied between  $vm \rightarrow$  and  $xm \rightarrow$ . Hence, more onlookers are recruited to richer sources and positive feedback behavior appears.

**Scout Bees Phase:** Scouts are the unemployed bees in the Artificial Bee Colony (ABC) algorithm that randomly choose their food sources. They play a crucial role in the algorithm as employed bees whose solutions cannot be improved through a predetermined number of trials, known as the "limit" or "abandonment criteria," become scouts. When an employed bee's solution is abandoned, the scout starts searching for new solutions randomly. This random search allows for the exploration of new areas in the search space. The abandonment and random search performed by scouts ensure that poor solutions, either initially or due to extensive exploitation, are discarded, creating a negative feedback behavior to balance the positive feedback in the algorithm.

The researchers decided to adopt Artificial Bee Colony (ABC) algorithm for classification of PDF documents, as it was found to be the most appropriate and convenient. The ABC algorithm is a nature-inspired optimization strategy that mimics honeybees' foraging behavior.

Among other reasons, one of the key benefits of using the ABC approach is its ability to exploit information in each individual PDF document. These could include titles, abstracts, keywords and full text itself among others. This ability enables extraction and application of such data by the researcher through ABC in categorization process.

Additionally, simplicity is another advantage of the ABC algorithm. It is rather simple in terms of both its implementation and design; lacking complicated setups or any need for extensive parameter controls. Consequently, this feature allows researchers to incorporate this algorithm into their text mining endeavors without experiencing significant computation overheads or technical complexities associated with it.

Finally, another factor that makes it favorable for use with several text mining and classification tasks is that ABC has been successfully applied. Document clustering, sentiment analysis as well as text categorization are some of the areas where it has shown potential results so far. The algorithm's ability to handle textual data effectively makes it a suitable choice for the researchers' task of categorizing PDF files.

## **Developmental Tools**

Developmental tools are vital for software development. They help developers create, test, and improve software. These tools offer various features like editing code, fixing errors, managing projects, and collaborating with others. They make software development easier and more efficient for developers.

### ***Python***

Python is a high-level programming language that was created by Guido van Rossum and first released in 1991. It is known for its simplicity and readability, making it a popular choice among programmers for a wide range of applications. Python emphasizes code readability by utilizing a clean and expressive syntax, which allows developers to write logical and concise code. Python supports multiple programming paradigms, including procedural, object-oriented,

and functional programming. It provides a large standard library that includes modules for tasks such as file I/O, networking, web development, and more, making it highly versatile (Python.org, n.d)

### ***Tkinter***

Python's Tkinter toolkit is a great library that provides a user-friendly interface. Python developers may find it very useful because they can easily incorporate a GUI to their applications. Tkinter has built-in widgets which assist programmers in creating windows, buttons, labels, text entry fields, menus and many more.

One of the main advantages of Tkinter is its simplicity in use. Its syntax is easy and understandable thus even beginners and those new to GUI development will not have difficulties working with it. This library helps create and manage GUI components concisely so that the drudgery from which the application developer goes through cannot be attributed to the coding but only to function design (Hanly & Koffman 2015).

Though this may imply that there are other graphical frameworks offering better visual appeal or advanced features than TKinter does; however, its ease-of-use, compatibility across platforms as well as its integration with Tcl/Tk has made it become one of the best choices for small-to-middle sized projects, quick prototyping and educational purposes (Python Software Foundation, 2021).

### ***Visual Studio Code***

Microsoft developed Visual Studio Code (VS Code), an open-source code editor that supports various programming languages and offers many features and extensions for a software

developer or programmer. There is no single document listing all VS Code features; however, its capacities have been exhaustively covered in official documentation, articles, tutorials among others.

The integrated version control system in VS Code includes Git integration which makes it easy for developers to manage their repositories without leaving their IDEs. Branch management, commit history, conflict resolution, among others are some of them which make use of Git commands within the editor itself leading to better workflow during version control (Visual Studio, 2015).

### ***Evaluation System ISO 25010***

Standard ISO/IEC 25010 is a quality model that comprehensively defines software product quality and it adds up to the fact that ISO/IEC 25010 is a standard that describes an integrated model for assessing software product quality.

This standard named “Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models” provides a systematic framework for evaluating the characteristics of software systems.

The eight main quality characteristics of the quality model described in ISO/IEC 25010 are functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability. Sub-characteristics are broken down into various aspects of software quality for each characteristic hence giving a precise breakdown of different elements contributing to these categories. This is necessary so that organizations can use established criteria to measure their systems against best practice to improve them accordingly so that they

meet acceptable standards on performance, safety or any other vital attribute among others as required by ISO/IEC/IEEE (2011).

### **Review of Related Studies**

Various studies have been carried out on the development of File Categorization using different approaches such as Rule-Based Approach, Machine Learning Approach, Neural Network-Based Approach and Hybrid Approach.

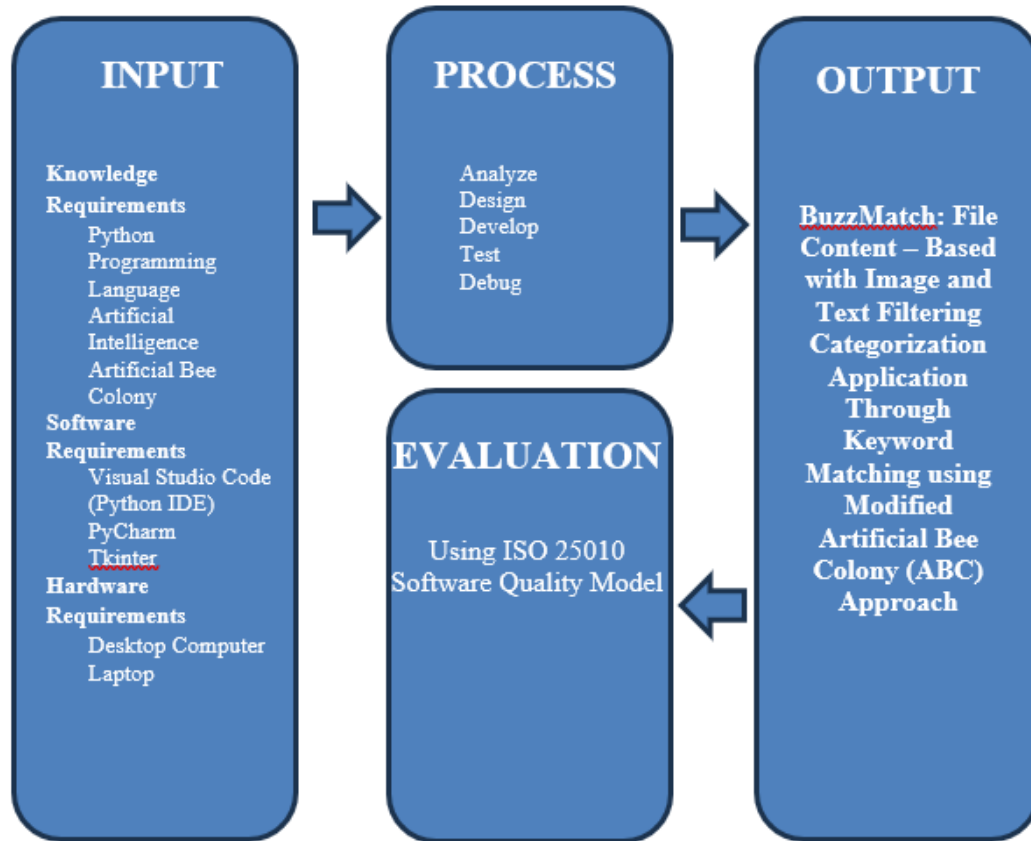
Many researchers have used this approach according to Baraka and Rezqa (2021) who started by extracting metadata information from documents. Metadata comprises document attributes such as author, title, creation date, and keywords associated with the document. Exploit this metadata for an understanding of what is in the documents and how they relate to each other. Secondly, these researchers use keyword extraction techniques. These are critical components of categorization that enable a user to understand what category or classification a given document falls into just by looking at its headings. The main goal is to summarize or describe these key features that show how the topic was symbolized in its statement.

In the fast evolution of artificial intelligence, other studies in relation to machine learning and neural networks have been done. One such research is focused on document classification via Support Vector Machine (SVM). A methodology is introduced that exploits SVM as a supervised machine learning technique for classifying documents according to their textual content. The method uses text-based features such as words and their counts as the input for training the SVM classifier. Consequently, this creates a feature space that represents these documents in terms of numbers capturing the distinguishing characteristics of different categories of documents. The paper provides an overview of the training process where an SVM

model learns to separate and classify documents by finding an optimal hyperplane in the feature space. It further emphasizes the importance of selecting appropriate kernel functions capable of handling nonlinear relationships and improving classification accuracy. Experimental evaluation using a widely diverse dataset shows that SVM outperforms other classification algorithms thereby indicating its efficacy and robustness in document classification tasks. From the study, SVM is one of those classifiers with better performance than others when it comes to categorizing documents based on textual features (Mayor & Pant 2012). Another study concentrated on document classification, and it has suggested a methodology using Naive Bayes and N-Gram. This is achieved by employing Naïve Bayes, which is a probability classifier, coupled with N-Gram models to assign documents depending on their textual content. N-gram models account for the presence of words or groups of words appearing together in documents resulting in significant contextual details enabling classification. The training process allows estimating posterior probabilities of belonging for different classes whenever a new instance becomes available. In this regard, labeled documents are used to train the Naïve Bayes classifier to estimate conditional probabilities of various classes. The model then determines the likelihood that a given text belongs to any class according to its n-gram features and chooses the most probable one among them. Therefore, experimental results have shown that this technique correctly categorizes texts upon the principle. What is significant about this research is that it demonstrates how useful Naive Bayes can be when combined with N-grams for document sorting (Mohamed 2015).



### Conceptual Model of the Study



**Figure 7.** Conceptual Model of the Study

### *Input*

The input includes knowledge requirements, software requirements, and hardware requirements. The study requires the knowledge of Python Programming as it will be used to develop the system, artificial intelligence as the swarm intelligence a subclass of artificial intelligence, and artificial bee colony, this will be the main algorithm that the researchers will use in developing the system. It also needs the software requirements this tool will be used to

apply the knowledge of the researchers IDE like visual studio code. In addition, frameworks like tkinter for building the GUI of the application.

### ***Process***

The software development lifecycle encompasses several stages, including analysis, design, development & debugging, and testing. During the analysis phase, requirements and objectives are identified. In the design phase, the software architecture and components are planned. Development & debugging involve coding, integration, and issue resolution. Finally, testing is conducted to verify functionality and performance. While it may not always be an entirely linear process, the stages in the SDLC exist to formalize and direct the systematic approach to development such that teams have a method to understand the requirements, turn those into workable solutions, put those solutions together with the fewest possible mistakes, and test the product to a point where people will want to use it.

### ***Output***

The output of the study: “BuzzMatch: File Content – Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach”.

### **Operational Definition of Terms**

The following terminologies are defined for a better understanding the study:

**Modified ABC (Artificial Bee Colony) algorithm** which is a customized version of the Artificial Bee Colony algorithm for improving file categorization by tuning its parameters and mechanisms.

**Categorization**, which is associated with the automatic classification of files into pre-defined groups by content analysis, using filter mechanisms of image and text, and keyword matching based on the modified Artificial Bee Colony.

**Content** refers to the information, material, or substance within a specific medium, e.g. txt, docx, pdf, and jpg.

**Filtering** either marking something allowable by, or disallowable to, the application on the basis of a set of criteria based off of key matching in the application itself.

**Keyword matching** is the process of recognizing, examining and comparing the exact words or terms in the files so as to link them with a relevant or predefined category.

**Portable Document Format (PDF)** is a file format used to present and exchange such kind of documents reliably, irrespective of software, hardware, and operating systems.

**TXT** is a plain text file format that does not contain any formatting and is universally supported by all platforms and applications.

## **Chapter 3**

### **METHODOLOGY**

This chapter includes the project design, development procedures, system operations, testing processes, and evaluation procedure of the system.

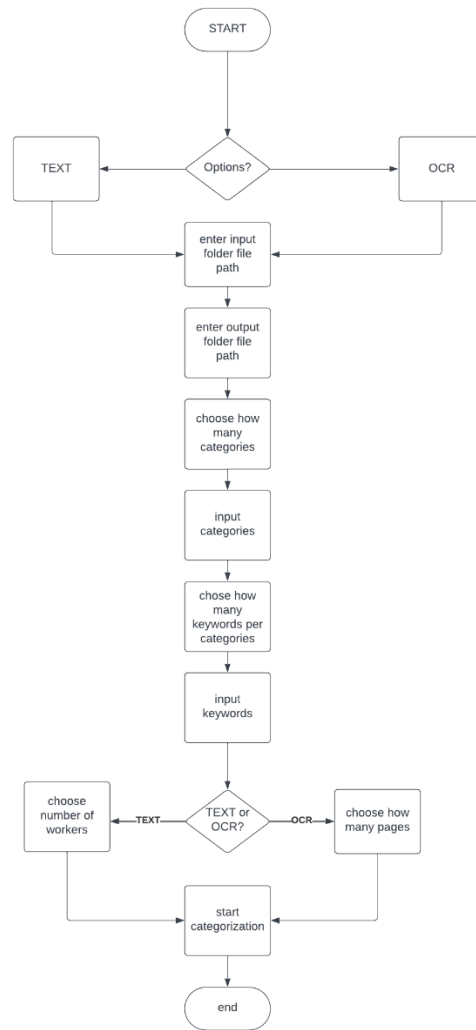
#### **Project Design**

The key purpose of this study is to come up with an automated software application that simplifies the classification of various document types based on their contents. The focused document formats include PDF, DOCX and TXT. In order to enrich the categorization, an Artificial Bee Colony (ABC) approach will be employed. ABC is a nature-inspired optimization algorithm which is expected to improve efficiency in terms of keyword matching and categorization process.

The proposed system design involves users inputting folder names and corresponding keywords for each folder. Subsequently, the system will employ parallel keyword matching, facilitated by the ABC algorithm, to simultaneously identify and categorize documents based on the specified keywords into their respective folders. The use of the Tkinter library in Python will provide a user-friendly graphical interface, ensuring ease of interaction with the software.

#### **System Design**

The general scope of this application is represented using the system flowchart as depicted in the figure below. The flowchart shows the structures that interact with the application as well as the flow of inputs and outputs.



**Figure 8.** System Flowchart

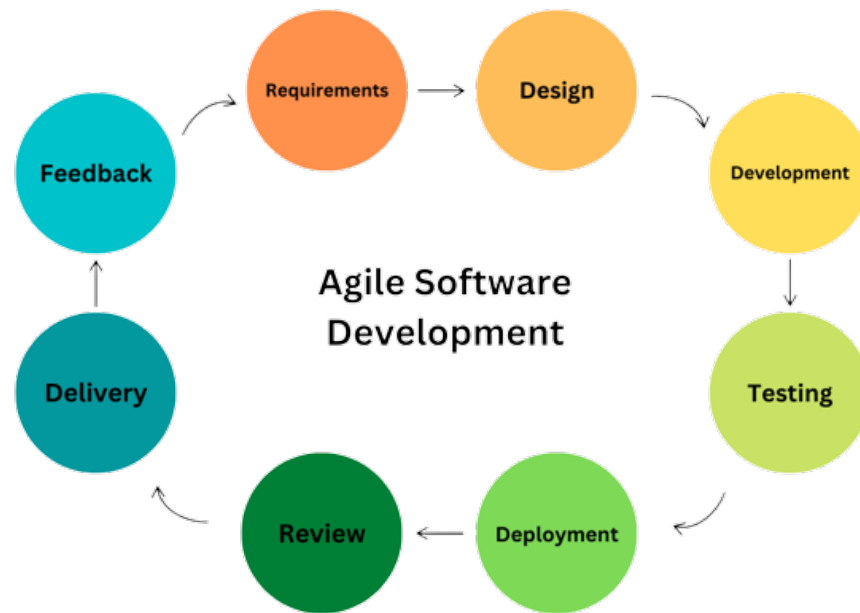
As for the system flowchart, it shows the decomposition of processes. This highlights the main functions of the application and its breakdown.

The initial step involves the user interacting with the application by selecting between text and OCR options. Upon selecting, the user is prompted to specify the input file path where the files requiring categorization are located, as well as the output file path where the categorized files will be stored. Subsequently, the user is required to determine the number of categories or folders needed and to provide the relevant keywords for each category. Depending on the user's

initial choice between text and OCR, the subsequent steps diverge slightly. For text input, the user will be prompted to specify the number of workers or CPU cores to be utilized for the categorization process. The categorization speed is directly influenced by the number of workers or cores specified. Once the user inputs this information, the categorization process commences. In the case of OCR, while the overall functionality remains similar to that of text input, a distinct step is introduced wherein the user must specify the number of pages to be analyzed by the OCR before categorization. This step ensures that the OCR processes only the required pages, thereby optimizing the categorization process. The differences between text and OCR processes primarily revolve around this additional step for OCR, emphasizing the selection of pages for analysis.

### **Project Development**

For system development, the researchers will use Agile software development methodologies to ensure a responsive and iterative approach to moving the project forward. Agile framework would help in ad-hoc changes as it will keep evolving and will refine the system based on continuous feedback.



**Figure 9.** Agile Software Development

**Requirements.** Before initiating any work, researchers need to compile a comprehensive list of user requirements for the mobile application, focusing on essential features. This involves prioritizing commonly used functionalities and deferring less frequently utilized ones for later consideration, allowing refinement after the initial release.

**Design.** Researchers in the design phase consider both interface and architecture of software. This is followed by selecting tools and frameworks to meet the requirements, creating a user interface (UI) mockup, which showcases prototype; this process also involves database design for efficient user data storage with great experience.

**Development.** In the development stage, the researchers participate in the system coding as designed in the many probable means and helps the project to set coding arrangement assist in

the refuse of computer application itertools described earlier. This is an obligatory stage as a lot depends on the groundwork of the project, and it undoubtedly requires a lot of time.

**Testing.** Functional Suitability (FST): Researchers runs a series of FSTs at this moment of time for the researchers and try to report the results before delivering the product to the client in order to show that various tests have an unfavorable result confirming that there are no mistakes or variances from the earlier changes to the product that would make it reliable.

**Deployment.** The testing is completed here and the application is uploaded on the server for beta testing or real use. This stage includes user training on how to use a system.

**Review:** This phase involves an appraisal by researchers on whether the application is market ready and coming up with answers for problems encountered during past rounds. Constant assessment and polishing form the foundation of this iterative notion.

**Delivery.** Once all preceding development phases are successfully navigated, it is time to deliver the application once again to the users.

**Feedback.** Using an evaluation tool, researchers collect user feedback during this stage to gauge the acceptance of the application. This valuable data is utilized to refine the application in subsequent iterations, ensuring continuous improvement.

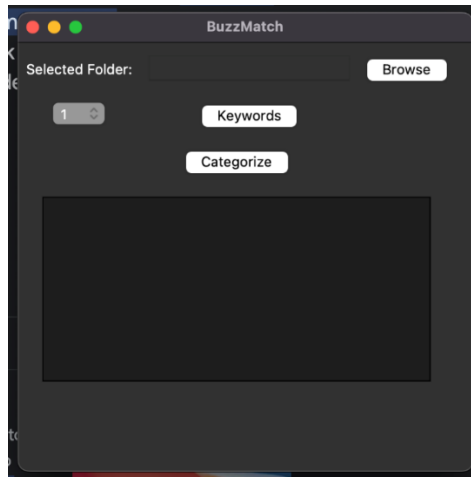
## **Operation and Testing Procedure**

The following procedure will be followed in order to operate the BuzzMatch File Content – Based document categorization application: as well as to test if all the intended system features are working properly, we will follow the some of the criteria in the ISO 25010 in creating a testing model.



## Testing Procedure

The test has been performed to check the features or functionalities of the software application. This is to ensure that each software component performs the expected output. The following steps had been taken for each iteration:



**Figure 10.** Sample UI for the Application

**Table 1.**

*Operating and Testing Procedure of the GUI Application (User Interface)*

| Test Module             | Steps To Undertake  | Expected Result  |
|-------------------------|---|--|
| Home Page Module        | <ol style="list-style-type: none"> <li>1. Select Text Categorizer.</li> <li>2. Select OCR Categorizer.</li> </ol>   | User should be able to select what type of categorizer the user will user.   |
| Text Categorizer Module | <ol style="list-style-type: none"> <li>1. Locate input file location.</li> <li>2. Locate output file location.</li> <li>3. Generate textbox.</li> <li>4. Generate keyword.</li> <li>5. Select how many</li> </ol> | <p>The user should be able to location input and output file location.</p> <p>The user should be able to choose how many textboxes and keyword the user wants.</p> |

---

|                        |   |       |  |
|------------------------|---|-------|--|
|                        | workers.<br>Press the<br>Categorization.  | Start | The user should be able to<br>categorize the pdfs and docx<br>files.   |
| OCR Categorizer Module | 1. Locate input file location.<br>2. Locate output file<br>location.<br>3. Generate textbox.<br>4. Generate keyword.<br>5. Select how many pages.<br>Press the Start Categorization |       | The user should be able to<br>location input and output file<br>location.<br>The user should be able to<br>choose how many textboxes<br>and keyword the user wants.<br>The user should be able to<br>categorize the pdfs and docx<br>files |

---

### Evaluation Procedure

To assess the acceptability of the developed application, 30 respondents will be conveniently selected, comprising 15 non-IT professionals, and 15 IT/CS professionals, using convenience sampling.

The application's features, both overall and within individual modules, will be presented to the respondents during a demonstration. Evaluators will then be invited to interact with the application as users, exploring its various features. Following the demonstration, evaluator-respondents will be provided with a questionnaire to gather their feedback. The collected questionnaires will process, and the data will be organized in an Excel file to calculate mean ratings.

The Likert Scale, as detailed in Table 2, was employed to interpret the adjectival ratings associated with the mean scores. The scale included categories such as "Highly Acceptable," "Very Acceptable," "Acceptable," and "Not Acceptable," with corresponding numerical ranges.

**Table 2.**

*Likert Scale*

| <b>Scale</b> | <b>Adjectival Rating</b> | <b>Range</b> |
|--------------|--------------------------|--------------|
| <b>4</b>     | Highly Acceptable        | 3.4 - 4.0    |
| <b>3</b>     | Very Acceptable          | 2.6 – 3.3    |
| <b>2</b>     | Acceptable               | 1.8 – 2.5    |
| <b>1</b>     | Not Acceptable           | 1.0 – 1.7    |

## Chapter 4

### RESULTS AND DISCUSSION

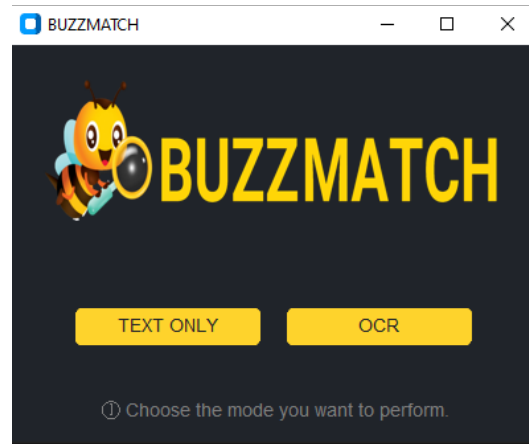
This chapter presents the results and discussion of the conducted study, it also includes the project description, project structure, and project capabilities and limitations, as well as the project test result and evaluation.

#### Project Description

The study developed a software application, 'BuzzMatch,' which automatically categorizes PDF, DOCX and TXT files using a multi-process approach inspired by the Artificial Bee Colony (ABC) algorithm. BuzzMatch leverages the core concept of parallel processing found in bee colonies to efficiently distribute the categorization task across multiple CPU core. BuzzMatch is designed to be a valuable tool for students, faculty, and professionals who need to efficiently organize and find documents.

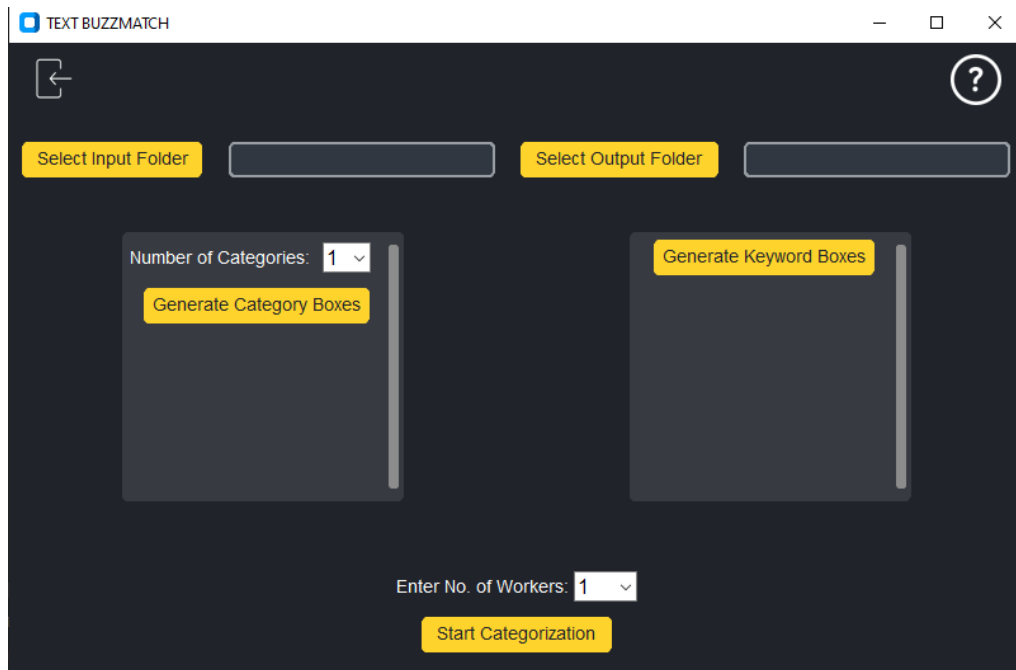
#### Project Structure

The computer application, which make use of artificial bee colony algorithm, will be used by the users. The application would focus on these modules: *Home Page Module*, *Text Categorizer Module* and *OCR Module*.



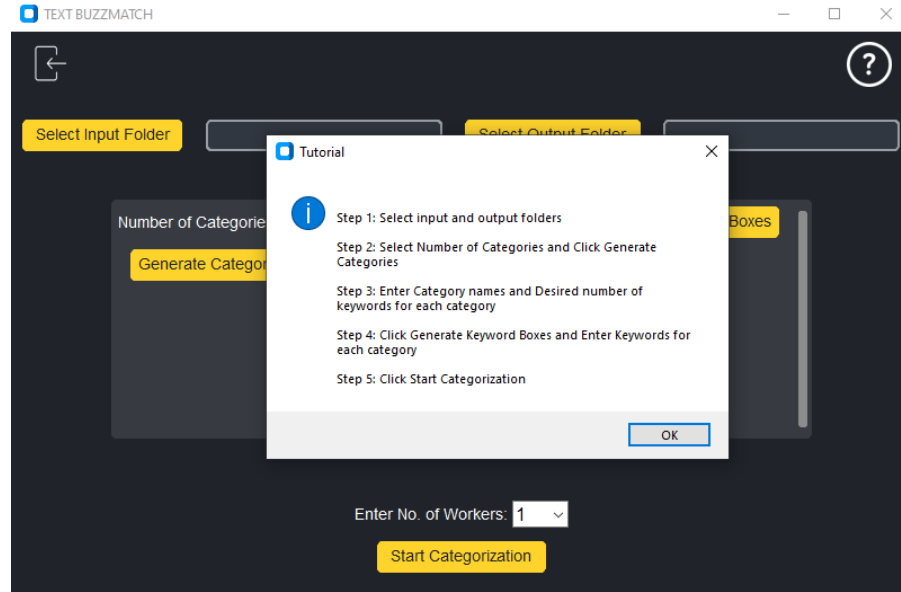
**Figure 11.** Home Page Module

The *Home Page Module*, as illustrated in Figure 11, serves as the gateway for users to select the type of categorization the user require. This module features two displayed buttons, each offering a distinct categorization option to meet user needs. The first button is Text Categorization, enabling users to categorize text-based content efficiently. This option is ideal for users who need to organize with pdfs and docx finding the corresponding keywords in the files. The second button provides access to Optical Character Recognition (OCR) Categorization, which is designed to handle the categorization of text extracted from images. This powerful feature is particularly useful for users who need to convert scanned documents, handwritten notes, or any other image-based text into a categorized format.



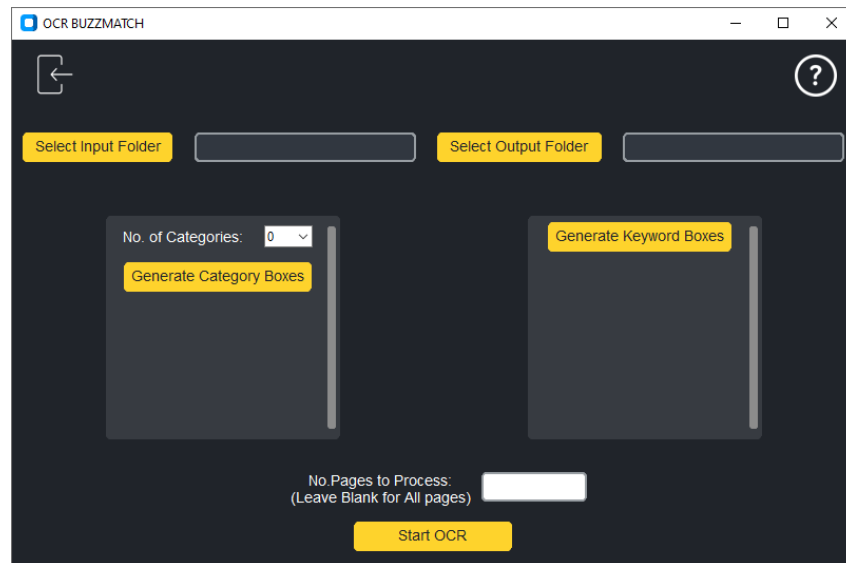
**Figure 12.** Text Categorizer Module

The *Text Categorizer Module* presented as shown in figure 12, contains information such as *select input folder* where in the user needs to locate the files the user wants to categorize and *select output folder*, where in the user needs to locate where the user wants to store the files in the input folder. In addition, there is also a button and dropdown on how many categories or folders the user wants to generate and generate keywords on how many keywords per categories the user wants to generate.



**Figure 13.** Instruction to use the Text Categorizer

The *Instruction* in *Text Categorizer Module* as shown in figure 13, shows the steps on how to use the Text Categorizer.



**Figure 14.** OCR Module

The *OCR Module* presented as shown in figure 14, contains information such as *select input folder* where in the user needs to locate the files the user wants to categorize and *select*

*output folder*, where in the user needs to locate where the user wants to store the files in the input folder. In addition, there is also a button and dropdown on how many categories or folders the user wants to generate and generate keywords on how many keywords per categories the user wants to generate.

## **Project Capabilities and Limitations**

The following are the capabilities of the system:

1. Dual Categorization Modes: Users can choose between Text Categorization (for TXT, PDFs and DOCX files) and Optical Character Recognition (OCR) Categorization (for image-based text).
2. Automated Categorization: The system automatically categorizes documents based on user-defined keywords and folders.
3. Customizable Categories: Users can create and name their own categories, specifying the number of keywords per category.
4. Error Handling: The application includes error handling mechanisms to manage incorrect or missing user input.

The following are the limitations of the system:

1. File Type Restrictions: The system is designed to handle only TXT, PDF and DOCX files. Other file types in the input folder will be moved to an "Uncategorized" folder in the output directory.



2. Keyword Matching: Documents that do not match any of the specified keywords will also be placed in the "Uncategorized" folder.
3. Keyword Logic: The keyword matching currently uses an "AND" logic, meaning a document must contain all keywords in a category to be assigned to that category.

## Test Results

The tables below show the testing procedure for each module containing the steps undertaken during testing and the actual results.

**Table 3.**

*Test Modules Result*

| Test Modules            | Steps Undertaken            |      |  | Test Results                                      |
|-------------------------|-----------------------------|------|--|---|
| Home Page Module        | 1. Select                   | Text |  | Users can choose what type of categorization.     |
|                         | 2. Select                   | OCR  |  |   |
| Text Categorizer Module | 1. Locate input             | file |  | Users can locate input and output file location.  |
|                         | 2. Locate output            | file |  | Users can generate how many textbox/folders.      |
|                         | 3. Generate textbox.        |      |  | Users can name the textbox for categories.        |
|                         | 4. Generate keyword.        |      |  |   |
|                         | 5. Select how many workers. |      |  | Users can put keywords per category.              |
|                         | 6. Press Categorization     |      |  | Users can select how many workers the users want. |
| OCR Categorizer Module  | 1. Locate input             | file |  | Users can locate input and output file location.  |
|                         | 2. Locate output            | file |  | Users can generate how many folders.              |

|                             |   |
|-----------------------------|---|
| 3. Generate textbox.        | many textbox/folders.                             |
| 4. Generate keyword.        | Users can name the textbox                        |
| 5. Select how many workers. | for categories.                                   |
| 6. Press Categorization     | Users can put keywords per category.              |
|                             | Users can select how many workers the users want. |
|                             | User can categorize txt, pdfs and docx files.     |

### Evaluation Result

A total of 30 respondents evaluated the system. Respondents were composed of 15 IT/CS professionals and 15 non-IT/CS professionals. Respondents evaluated the system using ISO 25010 with the following criteria: functionality, performance efficiency, usability, reliability, and maintainability.

**Table 4.**

*Responses to Functional Suitability*

|                            | Mean | Interpretation    |
|----------------------------|------|-------------------|
| Functional Completeness    | 3.5  | Highly Acceptable |
| Functional Correctness     | 3.63 | Highly Acceptable |
| Functional Appropriateness | 3.5  | Highly Acceptable |
| <b>Average</b>             | 3.54 | Highly Acceptable |

Under the *Functional Suitability* category, respondents evaluated the *Functional Completeness* of the system where 57% of the respondents (53% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 37% (47% of IT/CS professionals, and 27% of Non-IT/CS professionals) said that it is very acceptable and 6% (6% of Non-IT/CS

professionals) said that it is acceptable enough. Under the same category, respondents evaluated the *Functional Correctness* of the system where in the 70% (73% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 23% (27% of IT/CS professionals, and 20% of Non-IT/CS professionals) said that it is very acceptable and 6% (6% of Non-IT/CS professionals) said that it is acceptable enough. Moreover, the result for *Functional Appropriateness* of the system where 57% of the respondents (73% of IT/CS professionals, and 40% of Non-IT/CS professionals) said that it is highly acceptable and 37% (27% of IT/CS professionals, and 47% of Non-IT/CS professionals) said that it is very acceptable and 6% of the respondents (6% of Non-IT/CS professionals) said that it is acceptable enough.

**Table 5.**

*Responses to Performance Efficiency*

|                      | <b>Mean</b> | <b>Interpretation</b> |
|----------------------|-------------|-----------------------|
| Time Behavior        | 3.33        | Highly Acceptable     |
| Resource Utilization | 3.53        | Highly Acceptable     |
| <b>Average</b>       | 3.43        | Highly Acceptable     |

Under the *Performance Efficiency* category, the respondents evaluated the *Time Behavior* where the 47% of the respondents (47% of IT/CS professionals, and 47% of Non-IT/CS professionals) said that it is highly acceptable and 40% of the respondents (53% of IT/CS professionals, and 27% of Non-IT/CS professionals) said that it is very acceptable and 26% of respondents (26% of Non-IT/CS professionals) said that it is acceptable. In addition, the respondents evaluated the *Resource Utilization* where the 60% of the respondents (60% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 33% of the respondents (40% of IT/CS professionals, and 27% of Non-IT/CS professionals) said it is very acceptable and 7% of the respondents (7% of Non-IT/CS professionals) said that it is acceptable.

**Table 6.***Responses to Usability*

|                                 | Mean | Interpretation    |
|---------------------------------|------|-------------------|
| Appropriateness Recognizability | 3.6  | Highly Acceptable |
| Learnability                    | 3.57 | Highly Acceptable |
| Operability                     | 3.63 | Highly Acceptable |
| <b>Average</b>                  | 3.6  | Highly Acceptable |

Under the *Usability* category, respondents evaluated the systems *Appropriateness Recognizability* where 67% of respondents (73% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 27% of the respondents (27% of IT/CS professionals, and 27% of Non-IT/CS professionals) said that it is very acceptable and 6% of respondents (6% of Non-IT/CS professionals) said that it is acceptable. In addition, the respondents evaluated the *Learnability* where the 63% of the respondents (73% of IT/CS professionals, and 53% of Non-IT/CS professionals) said that it is highly acceptable and 30% of the respondents (27% of IT/CS professionals, and 33% of Non-IT/CS professionals) said it is very acceptable and 7% of the respondents (7% of Non-IT/CS professionals) said that it is acceptable. Moreover, the respondents evaluated the *Operability* where the 70% of the respondents (73% of IT/CS professionals, and 67% of Non-IT/CS professionals) said that it is highly acceptable and 23% of the respondents (27% of IT/CS professionals, and 20% of Non-IT/CS professionals) said it is very acceptable and 7% of the respondents (7% of Non-IT/CS professionals) said that it is acceptable.

**Table 7.***Responses to Reliability*

|              | Mean | Interpretation    |
|--------------|------|-------------------|
| Availability | 3.5  | Highly Acceptable |

|                 |      |                   |
|-----------------|------|-------------------|
| Fault Tolerance | 3.57 | Highly Acceptable |
| <b>Average</b>  | 3.54 | Highly Acceptable |

Under the *Reliability* category, the respondents evaluated the *Availability* where the 63% of the respondents (67% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 23% of the respondents (33% of IT/CS professionals, and 13 % of Non-IT/CS professionals) said that it is very acceptable and 14% of respondents (14% of Non-IT/CS professionals) said that it is acceptable. In addition, the respondents evaluated the *Fault Tolerance* where the 63% of the respondents (67% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 30% of the respondents (33% of IT/CS professionals, and 27% of Non-IT/CS professionals) said it is very acceptable and 7% of the respondents (7% of Non-IT/CS professionals) said that it is acceptable.

**Table 8.**  
*Responses to Maintainability*

|                | <b>Mean</b> | <b>Interpretation</b> |
|----------------|-------------|-----------------------|
| Reusability    | 3.6         | Highly Acceptable     |
| Analyzability  | 3.57        | Highly Acceptable     |
| <b>Average</b> | 3.59        | Highly Acceptable     |

Under the *Maintainability* category, the respondents evaluated the *Reusability* where the 57% of the respondents (60% of IT/CS professionals, and 53% of Non-IT/CS professionals) said that it is highly acceptable and 37% of the respondents (40% of IT/CS professionals, and 33% of Non-IT/CS professionals) said that it is very acceptable and 6% of respondents (6% of Non-IT/CS professionals) said that it is acceptable. In addition, the respondents evaluated the *Analyzability* where the 60% of the respondents (60% of IT/CS professionals, and 60% of Non-IT/CS professionals) said that it is highly acceptable and 27% of the respondents (40% of IT/CS professionals, and 13% of Non-IT/CS

professionals) said it is very acceptable and 13% of the respondents (13% of Non-IT/CS professionals) said that it is acceptable.

**Table 9.**

*Overall Summary of Responses*

|                        | <b>Total Mean</b> | <b>Interpretation</b> |
|------------------------|-------------------|-----------------------|
| Functional Suitability | 3.54              | Highly Acceptable     |
| Performance Efficiency | 3.43              | Highly Acceptable     |
| Usability              | 3.6               | Highly Acceptable     |
| Reliability            | 3.54              | Highly Acceptable     |
| Maintainability        | 3.54              | Highly Acceptable     |
| <b>Average</b>         | 3.53              | Highly Acceptable     |

The table 9 shows the Overall Summary of Responses in different evaluation. The table shows that the system obtained its highest rating under *Usability* with a weighted mean of 3.6 interpreted as “Highly Acceptable”. This rating indicates that the system was able to accurately process, and help the user to navigate the system, making it more user-friendly. Moreover, there are three categories, *Functional Suitability*, *Reliability* and *Maintainability* with the same results obtaining 3.54 weighted mean which is interpreted as “Highly Acceptable.” This implies that the system achieved the functions need and implied, a system that is operational and be able to maintain. The last one is *Performance Efficiency* obtaining a weighted mean of 3.43 interpreted as “Highly Acceptable”. This implies that the system achieved to save time, throughput parameters and is efficient in the use of resources under specified conditions.

## **Chapter 5**

### **SUMMARY OF FINDINGS, CONCLUSION, AND RECOMMENDATIONS**

This chapter recaps and concludes the results of the research from previous chapters. Included in this chapter is the summary of findings, final conclusions, and further recommendations for people who may want to pursue related studies to BuzzMatch.

#### **Summary of Findings**

Based on the results of testing procedure and evaluation, BuzzMatch was able to accomplish its objective and it can provide the users with tools they can use to categorize txt, pdfs and docx. The evaluation for the system received a favorable response for both IT/CS professionals and Non-IT/CS professionals, but still has room for improvement such as the user interface of the application, efficiency of the application and add more supported file formats.

#### **Conclusion**

The following conclusion were derived from the evaluation results of the developed automated categorizer using artificial bee colony algorithm:

1. Design a system with the following features:
  - a. Graphical User Interface (GUI):
    - i. Create an interactive UI for the users.
    - ii. Locate file destination.
    - iii. Ask for keywords.
  - b. Datasets
    - i. List of electronic text-based documents with random topics and contents
2. Create the system with developmental tools listed below:

- a. Front-end Tools
    - i. Python
    - ii. Tkinter
  - b. Back-end Tools
    - i. Python
3. Test and evaluate the system in terms of functional suitability and reliability.
  4. Determine the acceptability level of the developed system using the ISO 25010 criteria such as functional suitability, performance efficiency, usability, reliability, and maintainability.

## **Recommendation**

By implementing the recommended enhancements for the development of the application.

1. Enhance the application's user interface to increase its visual appeal and improve user engagement.
2. Incorporate a wider variety of text file formats. Expand the range to include additional types beyond the current selection.
3. Integrate the text categorizer system with the Optical Character Recognition (OCR) module for enhanced document analysis and categorization.
4. Use the evaluation to identify areas for further improvement in the scheduling system.
5. Make the system more specific use.



## APPENDIX A

## Survey Questionnaire Form

## BuzzMatch: File Content – Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified

Good Day!

We are fourth-year Computer Science students from the Technological University of the Philippines, and we are excited to present our thesis project, "**BuzzMatch: File Content-Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach**". Our application offers advanced image and text filtering capabilities for Word Documents (Docx) and Portable Document Formats (PDFs), utilizing a modified keyword matching algorithm.

We are reaching out to seek your valuable feedback and evaluation of our application. Your insights and expertise would be immensely helpful in refining and enhancing our project.

Thank you for considering our request. We look forward to your guidance and suggestions.

valloyasjonrexzel@gmail.com [Switch account](#)



Not shared

Next

Clear form

### Data Privacy Act of 2012

By participating in this survey, you consent to the following in accordance with the Data Privacy Act of 2012:

1. **Data Collection and Use:** Your personal information, responses, and any other data provided will be collected, processed, and used for the purposes of this survey. This includes analysis, reporting, and any related research activities.
2. **Data Storage and Security:** Your data will be securely stored and protected from unauthorized access, disclosure, or misuse. Appropriate security measures will be implemented to ensure the confidentiality and integrity of your information.
3. **Data Sharing and Disclosure:** Your data may be shared with third-party partners or service providers involved in the administration and analysis of the survey, but only to the extent necessary for these purposes. Your information will not be sold or shared with other third parties without your explicit consent.
4. **Anonymity and Confidentiality:** Efforts will be made to anonymize your data where possible. Any published results will not include personally identifiable information unless explicitly agreed upon by you.
5. **Rights to Access and Correction:** You have the right to access, correct, or delete your personal information at any time. Should you wish to exercise these rights, please contact us using the provided contact details.
6. **Withdrawal of Consent:** You may withdraw your consent and discontinue participation in the survey at any time without any negative consequences. To withdraw, please contact us, and your data will be promptly deleted.

By proceeding with this survey, you acknowledge that you have read, understood, and agreed to the terms outlined above. Thank you for your participation.

If you have any questions or concerns, please reach out to the following contacts:

### Demographic

What is your name? (Optional)

Your answer

What role do you take in this study? \*

☐ Professor

☐ IT Professional

☒ Student

☐ Other: \_\_\_\_\_

Back

Next

Clear form

### BuzzMatch: Instruction and Overview of the Application

The task of categorizing electronic text-based documents is challenging due to their varied and unstructured nature, making information extraction time-consuming and laborious. This issue is particularly pronounced with PDF files, whose global creation reached 2.5 trillion in 2020 (Rajeev, 2021), further complicating content management. Professionals and knowledge workers face significant inefficiencies, spending around 9.5 hours per week searching for documents with only a 50% success rate in retrieval (IDC, 2016). This problem hampers productivity and workflow optimization.

The increasing volume of electronic text-based documents combined with the time-consuming process of manually categorizing and retrieving them underscores the urgent need for efficient and automated methods of electronic text-based documents content categorization. Harnessing the power of artificial intelligence and machine learning algorithms may offer promising solutions to address this ongoing challenge, allowing users to streamline their document management processes and optimize their productivity.

Please watch the links for the overview and demo of the application.

Overview:

[BUZZMATCH OVERVIEW](#)

Demo:

[BUZZMATCH DEMO](#)

Bumalik

Susunod

I-clear ang form

### Functional Stability

This characteristic represents the degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions. This characteristic is composed of the following sub-characteristics:

★

|  | Not Acceptable<br>(1) | Acceptable (2)        | Very Acceptable<br>(3) | Highly<br>Acceptable (4) |
|--|-----------------------|-----------------------|------------------------|--------------------------|
| Functional Completeness -<br>Degree to which the set of functions covers all the specified tasks and intended users' objectives. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Functional Correctness -<br>Degree to which a product or system provides accurate results when used by intended users.           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Functional Appropriateness -<br>Degree to which the functions facilitate the accomplishment of specified tasks and objectives.   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |

Bumalik

Susunod

I-clear ang form

### Performance Efficiency

This characteristic represents the degree to which a product performs its functions within specified time and throughput parameters and is efficient in the use of resources (such as CPU, memory, storage, network devices, energy, materials...) under specified conditions. This characteristic is composed of the following sub-characteristics:

\*

|   | Not Acceptable<br>(1) | Acceptable (2)        | Very Acceptable<br>(3) | Highly<br>Acceptable (4) |
|---|-----------------------|-----------------------|------------------------|--------------------------|
| Time-behavior -<br>degree to which<br>the response and<br>processing times<br>and throughput<br>rates of a product<br>or system, when<br>performing its<br>functions, meet<br>requirements.   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Capacity: degree<br>to which the<br>maximum limits<br>of the product or<br>system, parameter<br>meet<br>requirements.   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Resource<br>utilization: degree<br>to which the<br>amounts and types<br>of resources used<br>by a product or<br>system, when<br>performing its<br>functions, meet<br>requirements. This<br>will primarily<br>depend on the<br>amount of files<br>and size of said<br>files. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |

## Usability

Degree to which a product or system can be interacted with by specified users to exchange information in the user interface to complete specific tasks in a variety of contexts of use. This characteristic is composed of the following sub-characteristics:

★

|  | Not Acceptable<br>(1) | Acceptable (2)        | Very Acceptable<br>(3) | Highly<br>Acceptable (4) |
|--|-----------------------|-----------------------|------------------------|--------------------------|
| Appropriateness<br>recognizability -<br>degree to which<br>users can<br>recognize whether<br>a product or<br>system is<br>appropriate for<br>their needs                   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Learnability -<br>degree to which a<br>product or system<br>enables the user to<br>learn how to use it<br>with effectiveness,<br>efficiency in<br>emergency<br>situations. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Operability -<br>degree to which a<br>product or system<br>is easy to operate,<br>control and<br>appropriate to use.   | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |

User error protection - degree to which a product or system protects users against making errors.

☐☐☐☐

User interface aesthetics - degree to which a user interface enables pleasing and satisfying interaction for the user.

☐☐☐☐

Accessibility - degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.

☐☐☐☐

## Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time. This characteristic is composed of the following sub-characteristics:

\*

|  | Not Acceptable<br>(1) | Acceptable (2)        | Very Acceptable<br>(3) | Highly<br>Acceptable (4) |
|--|-----------------------|-----------------------|------------------------|--------------------------|
| Availability -<br>Degree to which a<br>system, product or<br>component is<br>operational and<br>accessible when<br>required for use.                               | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| Fault tolerance -<br>Degree to which a<br>system, product or<br>component<br>operates as<br>intended despite<br>the presence of<br>hardware or<br>software faults. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |



## Maintainability

This characteristic represents the degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements. This characteristic is composed of the following sub-characteristics:

★

|  | Not Acceptable<br>(1) | Acceptable (2)        | Very Acceptable<br>(3) | Highly<br>Acceptable (4) |
|--|-----------------------|-----------------------|------------------------|--------------------------|
| <b>Reusability -</b><br>Degree to which a product can be used as an asset in more than one system, or in building other assets.  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| <b>Analysability -</b><br>Degree of effectiveness and efficiency with which it is possible to assess the impact on a product or system of an intended change to one or more of its parts, to diagnose a product for deficiencies or causes of failures, or to identify parts to be modified. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |
| <b>Testability -</b><br>Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/>  | <input type="radio"/>    |

**Comments and Recommendation**

Any comments, suggestions, and recommendations?

lyong sagot

**BuzzMatch: File Content – Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified**

Thank you for taking the time to participate in our thesis survey. Your valuable input is greatly appreciated and will contribute significantly to the success of our research. By sharing your insights and experiences, you are helping us gain a deeper understanding of our study topic, which will enhance the quality and impact of our findings. We are grateful for your support and collaboration in this important academic endeavor.

|  |  |                  |                     |
|--|--|------------------|---------------------|
| <br>VAA-COS | <b>TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES</b><br>Ayala Blvd., Ermita, Manila, 1000, Philippines<br>Tel No. +632-5301-3001 local 608   Fax No. +632-8521-4063<br>Email: cos@tup.edu.ph   Website: www.tup.edu.ph | Index No.        | REF-COS-3 5-INT-TGC |
|  |  | Revision No.     | 00                  |
|  |  | Effectivity Date | 06132022            |
|  |  | Page             | 1 / 1               |

## THESIS GRAMMARIAN CERTIFICATE

This is to certify that the thesis entitled,

**BUZZMATCH: FILE CONTENT – BASED WITH IMAGE AND TEXT FILTERING  
CATEGORIZATION APPLICATION THROUGH KEYWORD MATCHING USING MODIFIED  
ARTIFICIAL BEE COLONY (ABC) APPROACH**

authored by

**Espinosa, Joesef Andrei  
Riga, Rasheed  
Torres, Jan Christian  
Valloyas, Jon Rexzel**


has undergone editing and proofreading by the undersigned.

This Certification is being issued upon the request of Joesef Andrei Espinosa, Rasheed Riga, Jan Christian Torres, and Jon Rexzel Valloyas for whatever purposes it may serve them.

  
**DARREN JOE GACER FOLLERO**  
Grammarian

Date of Issuance

|                |                                   |
|----------------|-----------------------------------|
| Transaction ID | TUPM-COS-APS-WMM-01262024 -0244PM |
| Signature      |                                   |

|   |  |              |                 |
|---|--|--------------|-----------------|
|  | <b>TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES</b><br>Ayala Blvd., Ermita, Manila, 1000, Philippines<br>Tel No. +632-5301-3001 local 711   Fax No. +632-521-4063<br>Email: urds@tup.edu.ph   Website: www.tup.edu.ph | Index No.    | REF-URD-INT-CSI |
|   |  | Issue No.    | 01              |
| VRE-URD   | <b>CERTIFICATE OF SIMILARITY INDEX USING TURNITIN</b>  | Revision No. | 01              |
|   |  | Date         | 04132021        |
|   |  | Page         | 1 / 5           |
|   |  | QAC No.      | CC-04132021     |

This is to certify that the manuscript entitled

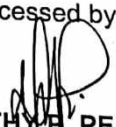
**"BUZZMATCH: FILE CONTENT-BASED WITH IMAGE AND TEXT FILTERING  
CATEGORIZATION APPLICATION THROUGH KEYWORD MATCHING USING  
MODIFIED ARTIFICIAL BEE COLONY (ABC) APPROACH"**

authored by

**JOSEF ANDREI ESPINOSA  
RASHEED RIGA  
JAN CHRISTIAN TORRES  
JON REXZEL VALLOYAS**

Has been subjected to similarity check on June 14, 2024 using Turnitin with  
generated similarity index of 14%.

Processed by:

  
**DOROTHY M. PERNIS**  
Staff, URDS

Certified correct by:

  
**FRANCISCO O. ESPONTILLA II, LPT, Ed.D.**  
Director, URDS

|                |  |
|----------------|--|
| Transaction ID |  |
| Signature      |  |

## References

Adobe Systems. (N.D.). *What is PDF?*. Retrieved from:

[https://www.adobe.com/ph\\_en/acrobat/about-adobe-pdf.html](https://www.adobe.com/ph_en/acrobat/about-adobe-pdf.html)

Ahlstrom, V (2005). *A COMPARISON OF SUBJECT-BASED CLASSIFICATION STRATEGIES*

*FORENHANCED USABILITY*. Retrieved from: [https://hf.tc.faa.gov/publications/2005-a-](https://hf.tc.faa.gov/publications/2005-a-comparison-of-subject-based-classification/full_text.pdf)

[comparison-of-subject-based-classification/full\\_text.pdf](https://hf.tc.faa.gov/publications/2005-a-comparison-of-subject-based-classification/full_text.pdf)

Amodeo, L., Lutz, F., Noubissi Tchoupo, M., Yalaoui, F. and Yalaoui, A. (2017). *Ant Colony Optimization Algorithm for Pickup and Delivery Problem with Time Windows*. Retrieved from:

[https://link.springer.com/chapter/10.1007/978-3-319-67308-0\\_19](https://link.springer.com/chapter/10.1007/978-3-319-67308-0_19)

Azizi, R. (2014). *Empirical Study of Artificial Fish Swarm Algorithm*. Retrieved from:

<https://arxiv.org/ftp/arxiv/papers/1405/1405.4138.pdf>

Baraka, R. S and Rezqa, E.Y. (2021) *Document Classification Based on Metadata and Keywords Extraction*, 2021 Palestinian International Conference on Information and

Communication Technology. Retrieved from: <https://ieeexplore.ieee.org/document/9637043>

Bigelow, S. J. (2023, June 5). *What is Open document format (ODF)?: Definition from*

*TechTarget*. Retrieved

from:

<https://www.techtarget.com/searchwindowsserver/definition/Open-Document-Format-ODF>

Blum, C. and Li, X. (2008). *Swarm intelligence in optimization*. In New optimization techniques

in engineering

(pp.

1-43).

Retrieved

from:

[https://www.researchgate.net/publication/227068137\\_Swarm\\_Intelligence\\_in\\_Optimization](https://www.researchgate.net/publication/227068137_Swarm_Intelligence_in_Optimization)

- Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From Natural to Artificial Systems*. Oxford University Press. Retrieved from: <https://www.orena/block/600989>
- Britannica, T. Editors of Encyclopaedia (2023, November 6). Microsoft Word. Encyclopedia Britannica. Retrieved from: <https://www.britannica.com/technology/Microsoft-Word>
- Buch, J. (2020). *PHISHING ATTEMPT OR NOT? HOW DIGITAL SIGNATURES IN PDF DOCUMENTS CAN HELP BRING TRUST TO YOUR BUSINESS*. Retrieved from: <https://www.entrust.com/blog/2020/06/how-digital-signatures-in-pdf-documents-can-help-bring-trust-to-your-business/>
- Cakir, A. (2016). *Usability and accessibility of portable document format*. Retrieved from: <https://www.tandfonline.com/doi/abs/10.1080/0144929X.2016.1159049>
- Çano, E., and Bojar, O. (2019). *Keyphrase generation: a multi-aspect survey*. Retrieved from: <https://ieeexplore.ieee.org/document/8981519>
- Choudhury, M., Counts, S., Gamon, M., Horvitz, E. (2021). . *Predicting depression via social media*. In Proceedings of the seventh international conference on Weblogs and social media (pp. 128-137). Retrieved from: <https://ojs.aaai.org/index.php/ICWSM/article/view/14432>
- Dorigo, M. and Strutzel, T. (2006). *Ant Colony Optimization*. Retrieved from: <https://web2.qatar.cmu.edu/~gdicaro/15382/additional/aco-book.pdf>
- Flynn, M. J. (1996). *Computer architecture: pipelined and parallel processor design*. Jones and Bartlett Publishers. Retrieved from: <http://arith.stanford.edu/solutions/solutions.pdf>
- Gelavska, A. (2023). *File Classification: A Complete Guide, Guidelines, and Process*. Retrieved from: <https://redfield.ai/file-classification/>

Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Professional, ISBN: 0201157675.

Government of Tamil Nadu (2019). *11<sup>th</sup> Office Management and Secretaryship*. Retrieved from: [https://www.brainkart.com/subject/Office-Management-and-Secretaryship-11th-std\\_347/](https://www.brainkart.com/subject/Office-Management-and-Secretaryship-11th-std_347/)

Gravett, A. (2017). *Ant Colony Optimisation-Based Algorithms for Optical Burst Switching Networks*. Retrieved from: <https://core.ac.uk/download/pdf/160256044.pdf>

IDC. (2016). The High Cost of Knowledge Worker Inefficiency: An IDC InfoBrief. Retrieved from <https://www.filestreamsystems.co.uk/the-high-cost-of-knowledge-worker-inefficiency-idc-infobrief/>

ISO/IEC/IEEE. (2011). Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. ISO/IEC 25010:2011. <https://iso25000.com/index.php/en/iso-25000-standards/iso-25010>

Jeanne, R. L. (1986). *The evolution of the organization of work in social insects*. Retrieved from:

Karaboga, D. (2005). *An idea based on honeybee swarm for numerical optimization*. Retrieved from: [https://abc.erciyes.edu.tr/pub/tr06\\_2005.pdf](https://abc.erciyes.edu.tr/pub/tr06_2005.pdf)

Kennedy, J. (2010). *Swarm Intelligence*. In Encyclopedia of Machine Learning. Retrieved from: <http://pzs.dstu.dp.ua/DataMining/bibl/Encyclopedia%20Machine%20Learning%202011.pdf>

Kennedy, J. and Eberhart, R. (1995). *Particle Swarm Optimization*. In Proceedings of IEEE International Conference on Neural Networks (pp. 1942-1948). IEEE.

Kim, E. (2022) *Importance of Records Classification & Tips to Improve Filing Accuracy*. . Retrieved from: <https://blog.collabware.com/2012/11/20/record-classification>

Latsoomanan, G. (2023). *Why the world needs to move away from papers and into digital documents*. Retrieved from: <https://www.linkedin.com/pulse/why-world-needs-move-away-from-papers-digital-gejamugan-latsoomanan>.

Loengarov, A. and Tereshko, V. (2005). *Collective Decision-Making in Honeybee Foraging Dynamics*. Retrieved from: [https://www.researchgate.net/publication/239463110\\_Collective\\_Decision-Making\\_in\\_Honey\\_Bee\\_Foraging\\_Dynamics](https://www.researchgate.net/publication/239463110_Collective_Decision-Making_in_Honey_Bee_Foraging_Dynamics)

Manning, C. D., Raghavan, P., & Schütze, H. (2008). . *Introduction to Information Retrieval*. Retrieved from: <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Martin J. and Shaw K. (2021). *10 top file-sharing services: Dropbox, Box, Google Drive, OneDrive, and more*. Retrieved from: <https://www.computerworld.com/article/3632983/top-file-sharing-services-dropbox-box-google-drive-onedrive-more.html>

Mayor, S. and Pant, B.(2012). *Document Classification Using Support Vector Machine*. Retrieved from: [https://www.researchgate.net/publication/266593700\\_Document\\_Classification\\_Using\\_Support\\_Vector\\_Machine](https://www.researchgate.net/publication/266593700_Document_Classification_Using_Support_Vector_Machine)

MESHDS. (2021). *The Disadvantages of Manual Document Filing Processes*. Retrieved from: <https://blog.mesltd.ca/the-disadvantages-of-manual-document-filing-processes-1>

MHC Team. (2021). *11 Benefits and Advantages of Document Management Systems*. Retrieved from: <https://www.mhcautomation.com/blog/11-benefits-and-advantages-of-document-management-systems/>



Microsoft. *Basic tasks in Word*. Retrieved from:

<https://support.microsoft.com/en-us/office/basic-tasks-in-word-87b3243c-b0bf-4a29-82aa-09a681999fdc>

Mohamed, K. (2015). *Using Naive Bayes and N-Gram for Document Classification*. Retrieved from: <https://www.diva-portal.org/smash/get/diva2:839705/FULLTEXT01.pdf>

Mu, J., Tang, H., Wang, M. and Wei, P. (2016). *An Improved Artificial Fish Swarm Algorithm and Its Application*. Retrieved from: [https://www.google.com/url?sa=t&ret=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiM4K7fiYmAAxWN9jgGHZ7iA\\_oQFnoECDQQAQ&url=https%3A%2F%2Fwww.atlantispress.com%2Farticle%2F25863600.pdf&usg=AOvVaw13cp37y0D1ia3Ky7gjo1Bp&opi=89978449](https://www.google.com/url?sa=t&ret=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiM4K7fiYmAAxWN9jgGHZ7iA_oQFnoECDQQAQ&url=https%3A%2F%2Fwww.atlantispress.com%2Farticle%2F25863600.pdf&usg=AOvVaw13cp37y0D1ia3Ky7gjo1Bp&opi=89978449)

Nguyen, H. (2023). *How To Search a PDF for Words or Phrases*. Retrieved from: <https://smallpdf.com/blog/how-to-search-a-pdf>

Nielsen, J. and Kaley, A. (2020). *PDF: Still Unfit for Human Consumption, 20 Years Later*. Retrieved from: <https://www.nngroup.com/articles/pdf-unfit-for-human-consumption/>

Python (N.D.). Tkinter. Retrieved from: <https://docs.python.org/3/library/tkinter.html>

Python.org. (N.D.). The Python Programming Language. Retrieved from: <https://www.python.org/>

Rajeev, R. (2021). *What Are The Main Reasons For The Increasing Popularity Of PDF Documents?*. Retrieved from: <https://www.managedoutsource.com/blog/what-are-reasons-for-increasing-popularity-of-pdf-documents/>

Rajeev, R. (2021). *What Are The Main Reasons For The Increasing Popularity Of PDF Documents?*. Retrieved from: <https://www.managedoutsourcing.com/blog/what-are-reasons-for-increasing-popularity-of-pdf-documents/>

Rosano. (2023). *8 Advantages of Using Digital Documents*. Retrieved from: <https://fintelite.ai/8-advantages-of-using-digital-documents/>

Rouse, M. (2016). *Text File*. Retrieved from: <https://www.techopedia.com/definition/9707/text-file>

Seeley, T. D. (1995). *The Wisdom of the Hive*. Retrieved from: <http://beekeeperstraining.com/file2/source/books/69.pdf>

Sheahan, K. (N.D) *File Organization Techniques*. Retrieved from: <https://smallbusiness.chron.com/organizational-skills-275.html>

Slettman, P. (2021). *Six Reasons To Digitize Important Documents*. Retrieved from: [www.forbes.com/sites/forbesbusinesscouncil/2021/12/08/six-reasons-to-digitize-important-documents/](http://www.forbes.com/sites/forbesbusinesscouncil/2021/12/08/six-reasons-to-digitize-important-documents/)

Tomokiyo, T., and Hurst, M. (2003). “*A language model approach to keyphrase extraction*,” in Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (Sapporo). Retrieved from: <https://dl.acm.org/doi/10.3115/1119282.1119287>

Visual Studio (2015), *Visual Studio Code*. Retrieved from: <https://code.visualstudio.com/docs>

Zhang, C. (2008). *Automatic keyword extraction from documents using conditional random fields*. J. Compute. Inform. Syst. 4, 1169–1180. Retrieved from: <http://eprints.rclis.org/12305/>

## RESEARCHER'S PROFILE

### Valloyas, Jon Rexzel D.

#117 Maria Clara St, 110, Caloocan City

+639614412039

[valloyasjr@gmail.com](mailto:valloyasjr@gmail.com)



## EDUCATIONAL BACKGROUND

### Tertiary

#### Technological University of the Philippines- Manila

Ayala Boulevard, Ermita, Manila

Bachelor of Science in Computer Science

2020 – Present

### Secondary

#### Columban College - Barretto

Olongapo City, Zambales

Senior High School

2018-2020

#### Barretto National High School

Olongapo City, Zambales

Junior High School

2014-2028

## SKILLS

**Programming Languages:** SQL, Python, C, C++, JavaScript, HTML, CSS, JS, SQL

**Specializations:** Web Development, Mobile Development, Database Management

## AFFILIATIONS

### Tertiary

#### Technical, TUPM - DOST Scholars' Club

Technological University of the Philippines –  
Manila

2023– Present

#### COS Representative, TUP Tech Guild

Technological University of the Philippines –  
Manila

2022– 2023

## PROJECTS

### **BuzzMatch: File Content-Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach**

Backend Developer

2023 - 2024

### **TUP Attendance System Mobile Application**

Mobile Developer

2023

## **Torres, Jan Christian L.**

26 C. Santiago St. Viente Reales Valenzuela City

09765488969

[janchristian.torres0709@gmail.com](mailto:janchristian.torres0709@gmail.com)



### **EDUCATIONAL BACKGROUND**

#### **Tertiary**

#### **Technological University of the Philippines- Manila**

Ayala Boulevard, Ermita, Manila

Bachelor of Science in Computer Science

2021-Present

#### **Secondary**

#### **Valenzuela City School of Mathematics and Science**

A. Pablo St, Valenzuela, 1400 Metro Manila

STEM Strand

2014-2021

### **SKILLS**

Web Development, Artificial Intelligence, Robotics Engineering, C, C++, C#, Python, HTML, CSS, JavaScript, Arduino

### **AFFILIATIONS**

#### **Secondary**

#### **President, Robotics Guild**

Valenzuela City School of Mathematics and  
Science 2014-2021

### **PROJECTS**

#### **BuzzMatch: File Content-Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach**

Thesis-UI/UX, Side-Features, Developer

January 2023-Present

#### **TUP Student Information System**

Web Developer

2023-2024

**iResearch: Valmasci's Research Papers Database**

Developer

2020

**Arthect: A WIFI-Based Household Electric Current Terminator**

Backend Developer

2021

**CanSatellite: DOST Project**

2019

**Espinosa, Joesef Andrei M.**

Block 25 Lot 12 Casimiro Townhomes Phase3 Pulang Lupa

Dos Las Pinas City

09156830548

[joesef07@gmail.com](mailto:joesef07@gmail.com)



## **EDUCATIONAL BACKGROUND**

### **Tertiary**

#### **Technological University of the Philippines- Manila**

Ayala Boulevard, Ermita, Manila

Bachelor of Science in Computer Science

2021-Present

### **Secondary**

#### **Holy Rosary Academy of Las Pinas City**

St. Joseph Avenue corner, Naga Road,

Pulanglupa Dos Las Pinas City, 1740

STEM Strand

2019-2021

#### **Young Achievers International School**

DBP Road, DBP Farms Subdivision, Las Piñas,  
Philippine

2014-2019

## **SKILLS**

C, C++, Python, , C#, .Net, OSI, TCP/IP, Linux, CloudFlare,DNS, Debian, FreeBSD, Kernel-  
Based Virtual Machines (KVM), Red Hat Enterprise Linux, TrueNAS, Hyper-V, VMWare,  
Visual Studio Code, Git, Github, Multimedia Software, (Photoshop, Gimp, Da vinci Resolve,  
Canva)

## **AFFILIATIONS**

### **Secondary**

#### **President, Supreme Student Government**

Holy Rosary Academy of Las Pinas City

2019-2020

#### **Writer, The Herald**

Young Achievers International School

2017-2018

## **PROJECTS**

### **BuzzMatch: File Content-Based with Image and Text Filtering Categorization Application Through Keyword Matching using Modified Artificial Bee Colony (ABC) Approach**

Thesis-UI/UX, Side-Features, Developer

January 2023-Present

### **TrueNAS Scale Homelab**

System Administrator

2022-Present

### **Pi-Hole DNS Sinkhole and Recursive DNS Server**

System Administrator

2022-Present

### **Age Detection and Protection of Minors in Media using Python and OpenCV**

Backend Developer

2023

### **Real Time File Integrity Monitor in PowerShell**

Developer

2023



## **Riga, Rasheed C.**

#38 Agueda st., Brgy. San Isidro, Cainta, Rizal

+639156823058

[rdriga7@gmail.com](mailto:rdriga7@gmail.com)



### **EDUCATIONAL BACKGROUND**

#### **Tertiary**

##### **Technological University of the Philippines- Manila**

Ayala Boulevard, Ermita, Manila

Bachelor of Science in Computer Science

2020 – Present

##### **Ateneo School of Government**

Ateneo de Manila University Katipunan

Avenue, Loyola Heights

Certificate Diploma in Financial Literacy and  
Social Entrepreneurship

2023-2024

#### **Secondary**

##### **Lyceum of the Philippines University – Manila**

Muralla St, Intramuros, Manila, 1002 Metro  
Manila

STEM Strand Specializing in Robotics

Engineering

2019-2020

##### **Philippine Normal University – Institute of Teaching and Learning**

Ayala Boulevard, Ermita, Manila

High School

2016-2019

### **SKILLS**

**Programming Languages:** SQL, Python, C, C++, PowerShell, Bash, Java, HTML, CSS, JS, R  
programming

**Specializations:** Artificial Intelligence, Data Visualization, Machine Learning, Database Management,  
Data Modelling, ETL and ELT, Data Science, Robotics Engineering, AI Robotics,

## **AFFILIATIONS**

### **Tertiary**

#### **Chief Technology Officer, Google Developers Student Club**

Technological University of the Philippines – Manila  
2024 – Present

#### **Artificial Intelligence/Machine Learning Core Lead, Google Developers Student Club**

Technological University of the Philippines  
2023-2024

### **Secondary**

#### **Member, Robotics Club**

Lyceum of the Philippines University-Manila  
2019-2020

#### **Member, Microsoft Student Ambassador**

Lyceum of the Philippines University-Manila  
2019-2020

## **PROJECTS**

**Fine-Tuned OpenAI Whisper Audio Transcriber in Tag-lish (Tagalog - English) Data using Google Fleurs Dataset**

**Motorcycle Helmet Detection Using YOLOv8**

**Person Detection Using YOLO**

BuzzMatch: File Content – Based with Image and  
Text Filtering Categorization Application Through Keyword Matching  
using Modified Artificial Bee Colony (ABC) Approach