Vol. No. I, July 2023

Mangalmay Institute of Engineering & Technology
Greater Noida, UP, India

Peer Reviewed

MIET - Journal of Engineers and Professional Management & Allied Research

Publisher Information

Publishing Body
Mangalmay Institute of Engineering & Technology (MIET)

Address
Plot no. 8, Knowledge Park-2, Greater Noida, UP 201310

Editor in Chief
Prof (Dr.) Yuvraj Bhatnagar (IQAC-Director), MIET, Greater Noida

Vol. I, July 2023 *Research Journal*

ISSN No……

Journal of Engineers and Professional Management & Applied Research

A Peer Reviewed Annual Interdisciplinary Research Journal

Mangalmay Institute of Engineering & Technology (MIET), Greater Noida

A Peer Reviewed Annual Interdisciplinary Research Journal of MIET, Greater Noida

Journal of Engineers and Professional Management & Applied Research

Published by: Mangalmay Institute of Engineering & Technology (MIET), Greater Noida

Mangalmay Institute of Engineering & Technology (MIET), Greater Noida

CONTENTS

A STUDY ON TECHNIQUES USING WINDOW KEYLOGGER IN THE PROFESSIONAL CAREER GROWTH

Aditya Rana[1], Kapish Kumar[2], Azeem Khan[3]

[1, 2, 3]Department of Computer Science,

MIET, Greater Noida, Uttar Pradesh, India

ABSTRACT

A keylogger, also known as a keystroke logger, is a software program or hardware device that records actions (keys pressed) on the keyboard. Keylogger is a type of spyware that keeps the user unaware that their actions are being tracked. Keyloggers can be used for a variety of purposes, including gaining illegal access to your private information by hackers and monitoring employee activities by employers. Some keyloggers, known as screen recorders, can also capture your screen at random intervals. Keylogger software records your keystrokes in small files that can be accessed later or emailed to someone watching your activity. Keyloggers are used in everything from Microsoft products or in any company's computers and servers. Sometimes someone can install/connect a keylogger to your phone or laptop to verify a fraud report. Worse, criminals have been known to infiltrate legitimate websites, apps, and even USB drives with keylogger malware. You should be aware of how keyloggers affect you, whether for malicious or legitimate purposes. Before delving into how keyloggers work, we'll first define keystroke logging.

Keywords: *Keylogger, Discord, Post-Exploitation*

INTRODUCTION

Keylogger is a tool/software that records everything on your computer or smartphone screen when you press a key on your keyboard or swipe the touchscreen. Keyloggers operate by listening in on your typing and recording it all into files that can later be accessed remotely via the Internet. Business owners use them to monitor productivity, so they know if their employees are doing their jobs correctly or not.

There are many different types of keyloggers available, each with a different price tag. Some keyloggers are small enough to hide on your machine yourself, or are placed on your desk to

hide all your typing and swiping movements.

Some keyloggers are installed on USB flash drives and other storage devices. Types of Keylogger

1) ***Hardware Keylogger****:* Hardware keyloggers are electronic devices designed to intercept data between a keyboard and an I/O port. These compact devices have an internal memory that stores keystrokes for later retrieval by the installer. Because they operate on the hardware platform, hardware keyloggers are undetectable by anti-viral software or scanners. Hardware keyloggers are far more powerful than software keyloggers, but they are less portable.

2) ***Software Keylogger****:* Keyloggers are activity-monitoring software programmes that allow hackers to access your data. When a user downloads an infected application, keylogger software is installed on the computer. It monitors the paths of the operating system that the keys you press on the keyboard must take once installed. This is how keylogger software tracks and records keystrokes. The information is then transmitted to the hacker via a remote server.

PROBLEM STATEMENT

Stealing user confidential data serves a variety of illegal purposes, including identity theft, banking and credit card fraud, and software andservice theft, to name a few. This is accomplished through keylogging, which can be considered eavesdropping, harvesting, and leakage of user-issued keystrokes. Keylogger is easy to install and use. When used for fraud purposes as part of more elaborate criminal heists, the financial loss can be significant. Signature-based solutions have limited applicability because they are easily evaded and also necessitate isolation. There is a lack of cyber security awareness among people regarding these attacks. Also, many keyloggerssoftware are paid.

ACTUAL METHODOLOGY

A keylogger is a type of malware or piece of hardware that monitors and records your The keylogger is designed to monitor people. It may also serve as parental supervision to monitor and prevent cyber incidents in kids. The purpose of this proposed research is to show how keyloggers log the victim's data and send it back to the attacker. we intend to keystrokes as you

type. It uses a command-andcontrol (C&C) server to send the information toa hacker.

A. **Requirement**s:

1) **Two machines**:

a) Victim machine

b) Attacker machine

2) **Tools required**: Discord.

3) **Software**: VMwareWorkstation

4) **IDE**: Sublime Text Editor

Research the effect of keyloggers on our system. In the proposed system, we use an executable file to capture the target's keystrokes and used Discord as a command-and-control server to send the data to the attacker.

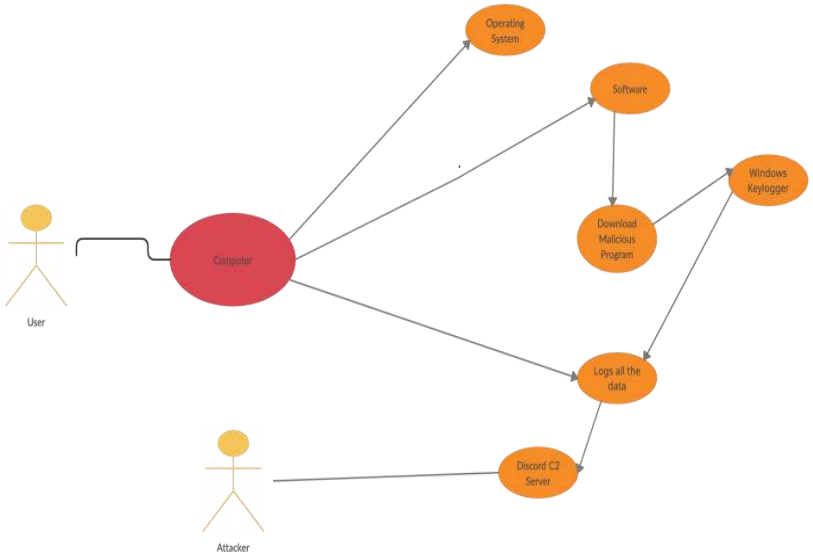


Figure 1: a flow diagram depicting the process of windows keylogger

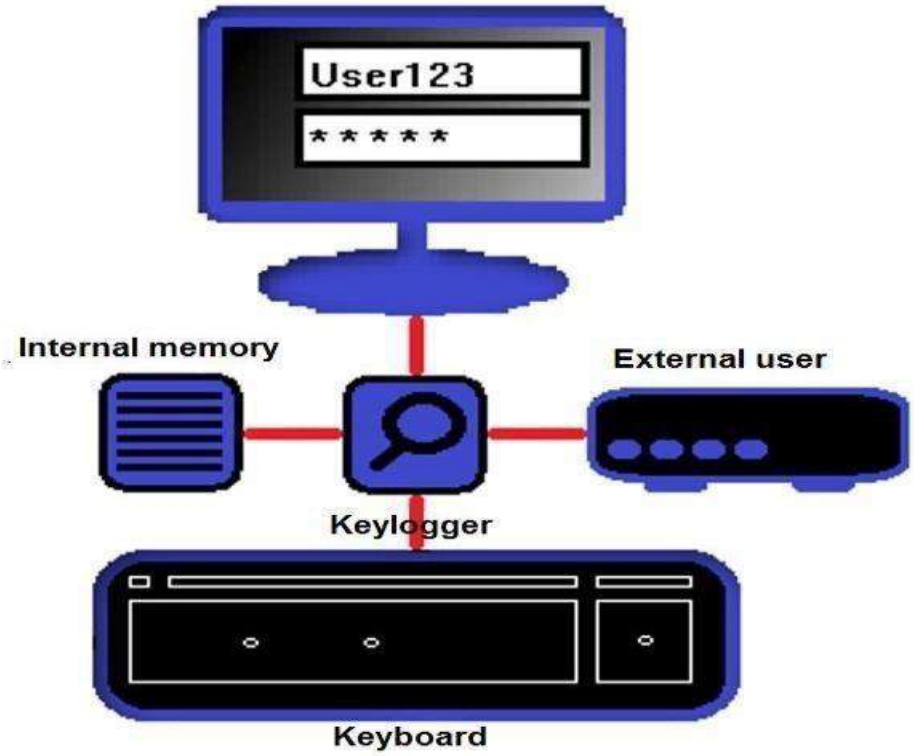


Figure 2: diagram of the log storage of keystrokes.

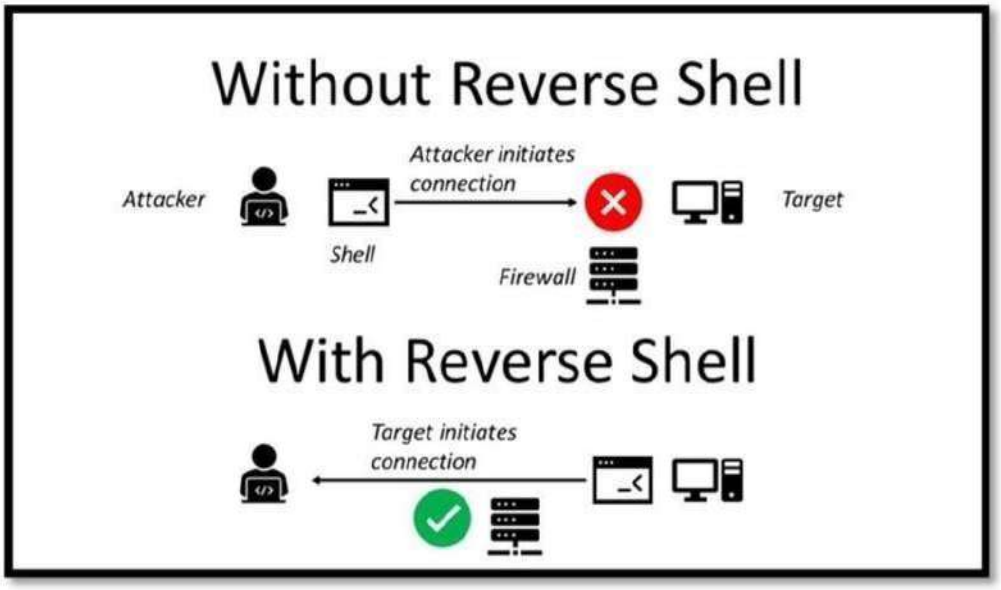


Figure 3: Reverse Shell

RESULTS

A. Attacker

Discord Server

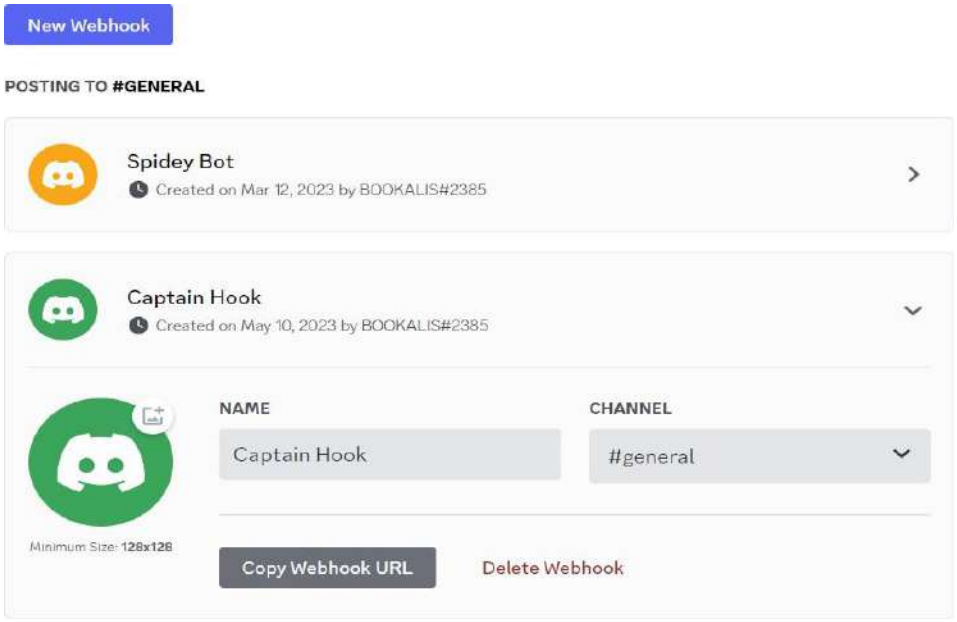


Figure 4: created a discord webhook to receive the keystroke data

In this, the attacker has created a discord server and webhook by which all the data can be received by an attacker on his c2 server.

B. Received Keystrokes of victim

Keystrokes of victim

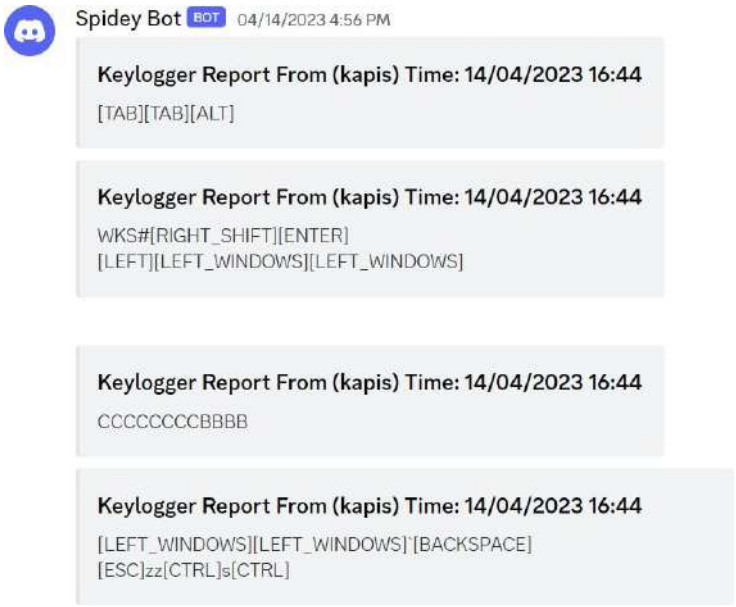


Figure 5: receiving the keystroke data into his control server.

In this, the attacker has received the keystrokes of the victim into his command-and-control server.

CONCLUSION

With the evolution of technology and the pervasiveness of computers in any private or industrial environment, keylogger devices, both hardware and software, pose a serious threat of cyber interception. Furthermore, because of the ease with which they can be found and purchased via the Internet at reasonable prices. The keylogger is a malicious programme that is difficult to detect and capable of reading and discovering anything on the keyboard. As a result, this survey paper is a comprehensive guide to everything you need to know about keylogger software. It's not always easy to tell if your device has a keylogger. In terms of hardware keyloggers, the only way to detect them is to inspect the keyboard, as well as the cables that connect to it.

REFERENCES

1 M. Aslam, R.N. Idrees, M.M. Baig, and M.A. Arshad. Anti-Hook Shield against the Software Key Loggers. In Proceedings of the 2004 National Conference on Emerging Technologies, pages 189–192, 2004.

2 Martin Vuagnoux and Sylvain Pasini. Compromising electromagnetic emanations of wired and wireless keyboards. In Proceedings of the 18th conference on USENIX security symposium, SSYM '09, pages 1–16, Berkeley, CA, USA, 2009. USENIX Association.

3 Mihai Christodorescu and Somesh Jha. Testing malware detectors. In Proceedings of the 2004 ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA '04, pages 34–44, New York, NY, USA, 2004. ACM

4 Manuel Egele, Theodoor Scholte, Engin Kirda, and Christopher Kruegel. A survey on automated dynamic malwareanalysis techniques and tools. ACM Computing Surveys (CSUR), 44(2):6:1– 6:42, March 2008. ISSN 0360-0300.

5 Andrea Lanzi, Davide Balzarotti, Christopher Kruegel, Mihai Christodorescu, and Engin Kirda. Accessminer: using system-centric models for malware protection. In Proceedings of the 17th ACM conference on Computer and communications security, CCS '10.

6 Kaspersky Lab. Key loggers: How they work and how to detect them. http://www.viruslist.com/en/analysis?pubid=204791931. Last accessed: Jan 2014.

7 Engin Kirda, Christopher Kruegel, Greg Banks, Giovanni Vigna, and Richard A. Kemmerer. Behavior-based spyware detection. In Proceedings of the 15th conference on USENIX Security Symposium, SSYM '06, Berkeley, CA, USA, 2006. USENIX Association.

8 Anthony Cozzie, Frank Stratton, Hui Xue, and Samuel T. King. Digging for data structures. In Proceedings of the 8th USENIX conference on Operating systems design and implementation, OSDI '08, pages 255– 266, Berkeley, CA, USA, 2008. USENIX Association.

9 Security Technology Ltd. Testing and reviews of key loggers, monitoring products and spy software. http://www.keylogger.org. Last accessed: Dec 2013.

AN EMPIRICAL STUDY ON PARADIGM SHIFT IN HUMAN RESOURCE RECRUITMENT AND MANAGEMENT SYSTEM IN THE CORPORATE

Harshita Singh[1], Deeksha Vishnoi[2], Aniket Dixit[3]

[1,2,3]Department of Computer Science & Engineering,

Mangalmay Institute of Engineering & Technology, Uttar Pradesh India.

ABSTRACT

The Human Resources Recruitment & Management System project is primarily focused with the administration of a company's Human Resource Department. The systems and processes at the confluence of human resource management and information technology are referred to as a Human Resource Management System (HRMS). It combines HRM as a discipline, and more specifically, basic HR activities and processes, with the information technology sector, whereas data processing system programming grew into standardized routines and packages of enterprise resource planning software. Acquiring and retaining high-quality talents is critical to an organization's success. As the job market becomes increasingly competitive and the available skills grow more diverse, recruiters need to be more selective in their choices, since poor recruiting decisions can produce long-term negative effects, among them high training and development costs to minimize the incidence of poor performance and high turnover which, in turn, impact staff morale, the production of high-quality goods and services and the retention the organizational integrity. At worst, the organization can fail to achieve its objectives thereby losing its competitive edge and its share of the market. The electronic human resource recruitment system has attracted a lot of attention because the human resource recruitment system has been a key point of enterprise recruitment and competition. This project argues the characteristics of traditional human resource recruitment and compares the types of human resource recruitment from five perspectives: quality, speed, dependability, flexibility, and cost. The process of onboarding employees, tracking their staffing capacity, and adding and terminating employees are all covered in this project. This project should keep track of each

employee and their company's staffing capacity, which can be used for performance evaluation. On this basis, transfers, removals, and promotions can be made.

Keywords: *Human Resource Management System (HRMS), HTML - hypertext markup language,CSS - cascading style sheets ,XML ,CSV - comma-separated values ,SQL - structured query language*

INTRODUCTION

In the 21st century, human beings are accelerating the speed to the information society, taking advantage of computer science is a long trend of enterprise recruitment in the future. Considering the development of technology and the importance of the human resource, managers realize that human resource is valuable, rare, inimitable and non-substitutable, electronic human resource recruitment has been a key point of the enterprise management and competition, so the electronic human resource recruitment system has attracted a number of attentions. Even though the human resource recruitment system is a large investment for organizations, it can save the cost for organization and increase the effectiveness.

The human resource recruitment system is a system to create a real-time, information-based, self-service, interactive work environment, and it refers to complete human resource recruitment practice with the advantage of computer-based technologies (Gainey, Klaas, 2003). Little attention was paid to exploring the differences between traditional human resource recruitment, and there was little literature showing how this project optimizes the human resource recruitment process. It is the reconstruction process of human resource recruitment. Before adopting the human resource recruitment system, managers should evaluate HR processes reengineering, this paper discusses the HR processes reengineering from the operation management perspective.

RELATED LITERATURE WORK

With the development of human resource recruitment system, comparing with the traditional human resource recruitment; it has several domains in this system: In the traditional human resource recruitment, the specialists are responsible for the different activities of HRM (e.g. Selection & Onboarding; Training & Development, Compensation Management, Employment Relationship), that means, the information of traditional human resource recruitment systems is separated in different units, the specialist may give feedback to employees about their

information, the information flow is sent from specialists to employees or their managers. But in the human resource recruitment system, all information is collected in the database; employees can get their own detailed information through their personal account in the system. For the specialists, they can acquire information from the database, and their information will be sent to the database. The data of the HRM is two-way flow (from specialist to database and from database to specialist). In traditional HR recruitment, the process of HRM is following those steps: (a) Sourcing (b) Screening, (c) Evaluation control, (d) Selection, (f) Employee relationship. Superior recruitment and selection strategies result in enhanced organizational results. With focus on this framework, the literature review of recruitment management systems will be prepared to shed light on Recruitment and Selection procedure. The core matter is to recognize universal practices which organizations adopt in recruitment and selection of employees then, to determine how the recruitment and selection procedures have effects on organizational results (Nel et al., 2004).

METHODOLOGY

Our application combines a number of inter-dependent methods of processing data. As a whole, our project uses Joomla for the backend and Application development technologies or the implementation of the recruitment management system. Each of the mentioned technologies is paramount to the creation of the application and the various steps it would take to produce the output. Web development methods are the primary part of the project as it is our output to the user, and will also be in use for the development of the end application. Application and Web Development Methods will also be in forefront to develop the product website and application. In order to achieve the stated objectives, the following methodology was used. Review of existing processes and systems to perform critical investigation and analysis of the existing recruitment process. System modeling using UML diagrams, use case and sequence diagrams to design/model the recruitment management system. Database management system (MYSQL) is used to create the database for the applicants and companies record.

Tools & Technologies used in Human Resources Recruitment & Management System

This section will describe various tools used to develop the project.

Joomla

Joomla is an open-source content management system used for creating Web content. It is written in PHP and makes use of a MySQL database for storing data and uses object-oriented programming techniques. It is one of the most popular content management systems owing to its features such as page caching, multi-language support, plugins and extensions.are created namely: res1 and res2, for storing crop names that satisfy preference conditions 10 For each candidate in recruitment dataset “candidates.csv ”do If preference 5 is satisfied then If preference 4 is satisfied then Append candidate name to list “res1 ”Else if preference 3 is satisfied then Append candidate name to list “res2 ”End if Else Reject the crop End if End for Finally, displaying predicted candidate list along with its details to the user.

MySQL

To connect python with the database, a MySQL connector is required. To work with the latest version it requires MySQL server version 8.0, 5.7, 5.6, 5.5. To install the MySQL connector, use command in the command prompt.

HTML

Hyper-Text Markup Language, regularly alluded to as HTML, is the standard markup dialect used to make website pages. Alongside CSS, and JavaScript, HTML is a foundation innovation used to make pages, and additionally to make user interfaces for portable and web applications.

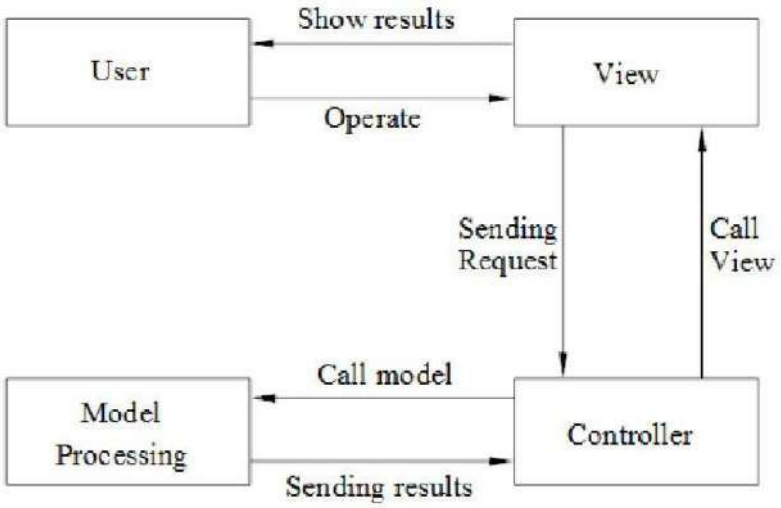
The following are basic hardware and software required to train and test the program.

PHP Version: Joomla recommends that you use a PHP Version that is either 5.6 and up or 7.0 and up.

PHPMyAdmin

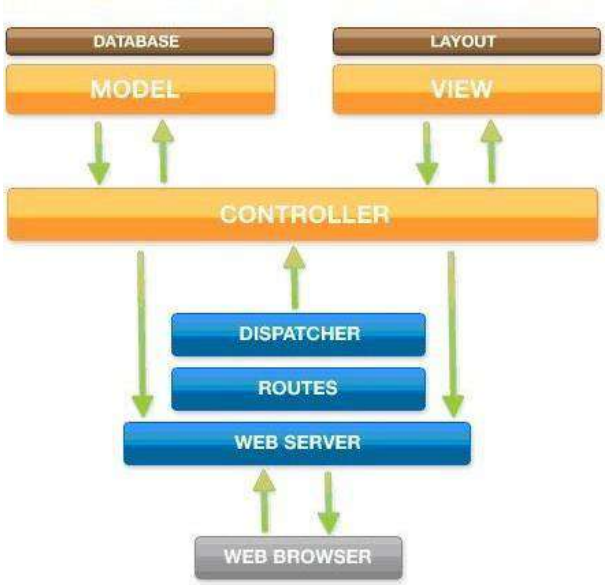
PhpMyAdmin, an apache server is required. provides both PHP- MyAdmin and apache for local computers. PhpMyAdmin is required to use the database for login and other purposes.

Data collection



Data Flow Diagram

The entity's data flows through multiple processes to produce results. The HRMS consists of model processing, which facilitates the discovery of information needs within the enterprise, and when a user or a manager requests access to the data, the call action will be performed with the help of the controller to generate the required data.



Design and architecture

Joomla Model View Controller (MVC)

Our application combines a number of inter-dependent methods of processing data. As a whole, our project uses Joomla for the backend and Application development technologies or the implementation of the recruitment management system. Each of the mentioned technologies is paramount to the creation of the application and the various steps it would take to produce the output. Web development methods are the primary part of the project as it is our output to the user, and will also be in use for the development of the end application. Application and Web Development Methods will also be in forefront to develop the product website and application. In order to achieve the stated objectives, the following methodology was used. Review of existing processes and systems to perform critical investigation and analysis of the existing recruitment process. System modeling using UML diagrams, use case and sequence diagrams to design/model the recruitment management system. Database management system (MYSQL) is used to create the database for the applicants and companies record.

User Authentication use case Narrative

This section displays the user authentication use case narrative alongside the use case diagram

Brief Description	This singular module is for getting access into the system by the user.
Actor(s)	Users
Flow Of Events	**Basic Flow:** The use case begins when the user accesses the webpage 1. The user enters the URL to the page. 2. The user inputs his or her login details. 3. System Displays Homepage **Alternative Flow:** If user information is incorrect he or she is not granted access to the recruitment platform.
Level	User use case
Parameters	**Input:** user login details **Output:** The recruitment platform homepage
Pre-Conditions	All users must: • Have valid user account. • Have working Internet connection.
Post-Conditions (Success End)	If use case is successful, user is granted access to the System.

for the user authentication module.

Employer use case narrative

Brief Description	This module gives the job provider the ability to view his or her information and also to edit it.
Actor(s)	Job providers
Flow Of Events	**Basic Flow:** The use case begins when the user has successfully logged in to the system. • The Job provider is directed to the homepage from the login page where he/she can decide to direct to his/her information page. • The job provider can decide to update the information which he/she initially inputted when he/she registered. • System displays job seeker information. • Job provider can decide to post a job opening. **Alternative Flow:** If the job provider is not logged in he or she is redirected to the login page.
Level	User use case
Parameters	**Input:** session values **Output:** the job provider information page.
Pre-Conditions	The job provider must have been successfully authenticated.
Post-Conditions (Success End)	If use case is successful, the job provider information page is displayed.
Post-Conditions (Failed End)	If use case is not successful, an error page is returned or in some cases the user is redirected to the login page.
Trigger	Webpage getting the correct session values

This section displays the user authentication use case narrative alongside the use case diagram for the user authentication module.

RESULT AND DISCUSSION

One of the fundamental principles of strategic HRM research is that the impact of HR practices on individuals as well as organizations is best understood by examining the system of HR practices in place. Considering that HR practices are rarely, if ever, used in isolation, failure to consider all of the HR practices that are in use to manage employee's neglects potential important explanatory value of unmeasured HR practices. Yet, while researchers may agree that a systems perspective is most appropriate, adopting a systems perspective introduces a host of issues and problems that remain to be addressed in literature. Our aim of this work is to develop a human resources recruitment management system.The following are the objectives that will be used to achieve this aim:

- To design/model the recruitment management system
- To create a database system for the applicants and companies record
- To implement the recruitment management system.

In the 21st century, human beings are accelerating the speed to the information society, taking advantage of computer science is a long trend of enterprise recruitment in the future. Considering the development of technology and the importance of the human resource, managers realize that human resource is valuable, rare, inimitable and non-substitutable, electronic human resource recruitment has been a key point of the enterprise management and competition, so the electronic human resource recruitment system has attracted a number of attentions. Even though the human resource recruitment system is a large investment for organizations, it can save the cost for organization and increase the effectiveness.

Human Resources Recruitment & Management system, is a time and money-saving platform that businesses may utilize to lower their recruitment costs. The burden of employers can be eliminated and reduced to a minimum with this research effort. The goal of this project is to make the onboarding process simple and efficient. E-recruitment can be improved even more by adding modules or functions that allow candidates to be tested and then referred to employers based on their scores and rankings on the profiles in order to have a real-time track over the employee.

REFERENCES

1. Evaluation and analysis of strategic human resource management based on multi-mode fuzzy logic control algorithms by Feng Jin;Li Wang - 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE).
2. Improvement In Manpower Productivity By Using Training Within Industry Job Methods (JM) (A Case Study Of Parason Group, India) Meenakshi Tyagi; Shivani Agarwal; Gagneet Kaur Bhatia 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).
3. Impact of Human Resource Practices on Knowledge Management: An Empirical Analysis Rajni 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
4. Agarwala, T. (2003). Innovative human resource practices and organizational commitment: An empirical investigation. International Journal of Human Resource Management, 14, 175–197. Ahmad, O., & Schroeder, R. G. (2003). The impact of human resource management practices on operational performance: Recognizing country and industry differences. Journal of Operations Management, 21, 19–43.
5. Allen, D. G., Shore, L. M., & Griffeth, R. W. (2003). The role of perceived organizational support and supportive human resource practices in the turnover process. Journal of Management, 29, 99–118.
6. Huselid, M. A., & Becker, B. E. (2000). Comment on ‘‘measurement error in research on human resources and firm performance: How much error is there and how does it influence effect size estimates?’’ by Gerhart, Wright, McMahan, and Snell. Personnel Psychology, 53, 835–854.
7. Ichniowski, C., Shaw, K., & Prennushi, G. (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. The American Economic Review, 87, 291–313.
8. Jackson, S. E., Chuang, C., Harden, E. E., & Jiang, Y. (2006). Toward developing human resource management systems for knowledge-intensive teamwork. In: J. Martocchio (Ed.), Research in personnel and human resource management. Jackson, S. E., & Schuler,

R. S. (2000). Managing human resources: A partnership perspective. Cincinnati, OH: South-Western. Jackson, S. E., Schuler, R. S., & Rivero, J. (1989). Organizational characteristics as predictors of personnel practices. Personnel Psychology, 42, 727–786.

9. James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. Psychological Bulletin, 81, 1096–1112. Johnson, J. W. (1996). Linking employee perceptions of service climate to customer satisfaction. Personnel Psychology, 49, 831–851.

10. Katz, H. C., Kochan, T. A., & Keefe, J. H. (1987). Industrial relations and productivity in the U.S. automobile industry. Brookings papers on economic activity. Washington: The Brookings Institute. Klein, K. J., & Sorra, J. S. (1996). The challenge of innovation and implementation. Academy of Management Review, 21, 1055–1088.

11. Koch, M. J., & McGrath, R. G. (1996). Improving labor productivity: Human resource management policies do matter. Strategic Management Journal, 17, 335–354.

A DIAGNOSTIC ANALYTICAL EXPLORATION ON POTATO LEAF DISEASES AND ITS CLASSIFICATION USING CNN

Mohit Ranjan[1], Mohan Joshi[2], Ghanshyam Yadav[3]

[1,2,3]Deparment of Computer Science Engineering,

Mangalmay Institute of Engineering and Technology, Greater Noida, UP, India

ABSTRACT

Potato is one of the staple foods that widely consumed, becoming the 4th staple food consumed throughout the world. Also, the world demand for potato is increasing significantly, primarily due to the world pandemic corona virus. However, potato diseases are the leading cause of the decline in the quality and quantity of the harvest. Inappropriate classification and late detection of the disease's type will drastically worsen the plant conditions. Fortunately, several diseases in potato plants can be identified based on leaf conditions. Therefore, in this paper, we present a system to classify the four types of diseases in potato plants based on leaf conditions by utilizing deep learning using the VGG16 and VGG19 convolution neural network architecture model to obtain an accurate classification system. This experiment has achieved an average accuracy of 91-93%, which indicates the feasibility of the deep neural network approach.

Keywords: *Leaf Disease Classification, Deep Learning, VGG16, VGG19, Potato Plant*

INTRODUCTION

The agricultural sector is one of the essential role holders in Indonesia, according to a survey conducted by the Indonesian statistics show that 28.79% of Indonesia's population depend on agriculture [1] [2]. Food safety and nutrition improvement are some of the significant challenges faced by the agricultural sector. Potatoes become one of the staple foods that are expected to be able to suffice these needs in terms of quantity and quality. They are rich in nutrients, most notably vitamins C and B6 and the minerals, potassium, magnesium, and iron [3]. Besides being a popular staple food in Indonesia, potatoes are the number one vegetable crop in the United States of America and became the fourth most consumed vegetable crop in the world. Potato agricultural products have developed rapidly in this decade. Every year, the amount of production can reach around 850,000 tons. The amount is produced from an area of about 60,000

hectares. The area of planting and production has increased by approximately 10% per year, making Indonesia the largest potato producing country in Southeast Asia. The presence of pests and diseases of potato plants during the growth period has reduced the quality and quantity of agricultural products. Potato pests and diseases can lead to an early harvest where the harvest is done when the potatoes are still small in size, and crop failure is caused by spoilage of the potato plants before harvest. These problems are mostly caused by the late identification of diseases in potato plants and errors in disease diagnosis.

Fig. 1. Input Image of Potato Leaf Disease Classification.

Fig. 2. Output Image of Potato Leaf Disease Classification.

The identification of diseases in potato plants quickly and accurately is highly essential to reduce the impact of diseases on plants. Manual monitoring activities carried out by farmers become difficult and impractical because it takes a long time and in-depth knowledge. Identification of plants diseases types that are slow will trigger the spread of diseases in plants uncontrollably. Besides, farmers generally identify diseases in plants in a way that is approximately and assumptions that allow inaccurate identification results because the symptoms on the leaves appear to have similarities that are difficult to describe at a glance. Farmers use the results of personal identification without expert advice in the field of plant diseases as a reference in the prevention of plants infected with the disease. As a result, preventive measures taken by farmers may be ineffective and can damage crops due to inadequate knowledge and misinterpretation of disease intensity, excessive dosage, or lack of dosage [4]. This problem is the foundation of the

proposed research to facilitate farmers in identifying and classifying diseases in potato plants that are fast and accurate.

The proposed research methodology focuses on the classification and identification of healthy and disease- infected leaf conditions using the Deep Learning approach as in Fig. 2. The architecture used in this study is VGG16 and VGG19, which is a Convolutional Neural Network architecture model of the VGGNet Group. Both architectures have five layers of convolution and have a difference in the number of layers. In accordance with the names of the two, VGG16 has 16 layers and VGG19 has 19 layers. The architectural layer allows us to study effective features for classifying diseases through leaves.

RELATED WORK

Research in the field of disease classification in plants has been carried out even with various methods. However, it is considered still lacking [5] and become one area of research that continues to be developed because the subject in this field varies significantly. Thirty-seven papers in agriculture were published in the 2015-2017 period. More precisely, 15 papers were published in 2017, 15 in 2016, and 7 in 2016. This fact shows how new and modern this technique is in the agricultural domain [6].

In order to learn an effective features representation, a deep learning-based method can be devised. Deep learning [7] has shown very good performance in many visual perception tasks, such as text detection [8], [9], victims detection, [10], [11], target tracking [12], [13], object detection [14], [15], [16]. A deeper network can boost accuracy. Interestingly, the work [10] has been proven theoretically and empirically that the last layers of the deep network can capture more semantics information or abstraction; thus, it is more robust to variation of pose, colour, scale, and deformable object, that is it could be suitable to classify the leaf diseases robustly.

Prajwala et al. [4] This paper uses the LeNet architectural model, Convolutional Neural Network model architecture to detect and identify tomato leaf diseases that are commonly found in tomato additions; Septoria Leaf Spot and Yellow Leaf Curl. The dataset was obtained by researchers from one of the Open Access image databases, PlantVillage. The level of accuracy resulting from the methodology proposed for this paper is 94% -95%.

Experimental results on a developed model by Srdjan Sladojevic et al. [17] with a new approach to recognize diseases in five types of plants and 13 different kinds of diseases using the convolutional neural network (CNN) method. From this study, an average accuracy of 96.3% was obtained. Using the same method but with a different architecture, Erika Fujita et al. [18] made a disease diagnosis system in cucumber plants using the Convolutional Neural Network method by adopting the AlexNet architecture and obtaining an average accuracy of 82.3%.

A total of 2000 images of corn leaf were obtained by Raja P et al. [19] from the open-access image database, PlantVillage. The dataset is used as an object classification of 3 types of diseases in maize leaves using a bag of features with the Multiclass SVM (Support Vector Machine) method. On this occasion, the researchers also evaluated the accuracy of the method compared to the histogram method and feature-based grey level co-occurrence. The results obtained show that the SVM (Support Vector Machine) Multiclass method is entirely accurate. Pranjali B. Padol and Prof. Anjali A. Yadav [20] uses segmentation with K-mean clustering to found the disease region, and then the colour and texture of the image are extracted. While for classification, they use the SVM (Support Vector Machine) Classifier method and get an accuracy of 88.89%.

Jobin Francis et al. [21] was adopting the Backpropagation Neural Network method to classify disease types in pepper plants and GLCM (Gray Level Co-occurrence) for feature extraction step. Types of pepper plant diseases that can be detected are Berry Spot Disease, a kind of fungal infection found in pepper and Rapid Disease, a kind of disease caused by mineral deficiencies such as nitrogen, magnesium, and potassium. Eftekhar Hossain et al. [22] used the same method for the classification of diseases such as Alternaria Alternata, Anthracnose, Bacterial Rot, Leaf Spot, and Plant Leaf Cancer. The performance of the KNN disease detection system in this paper is 96.76% accuracy.

Aakanksha Rastogi et al. [23] conducted a study using artificial neural networks and fuzzy logic methods to detect diseases in plants based on leaf conditions. This study focuses on identifying and classifying diseases in maple and hydrangea leaves. The condition of leaves of plants affected by the disease is divided into two categories, namely leaf spot, and scorch leaf. Leaf spots are where the disease on the leaves is at specific points of the leaf, while scorched leaves are where the disease on the leaves spreads evenly on the leaves.

PROPOSED METHODOLOGY

As shown in Fig. 3. The proposed methodology in this paper includes the following four main steps: data acquisition, data pre-processing, data augmentation, and image classification.

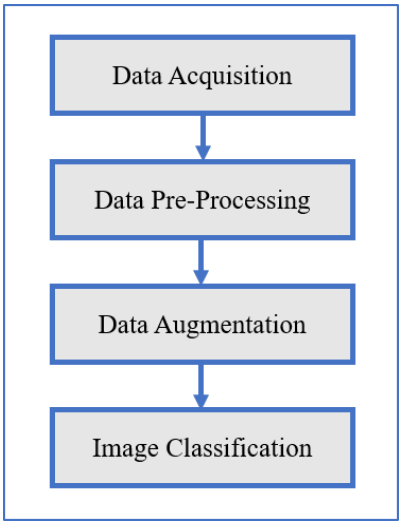


Fig. 3. Diagram Block Proposed Methodology.

A. Data Acquisition

Different image resolutions and sizes were obtained from several sources, including those collected by authors from a potato plantation in Malang, Indonesia, PlantVillage [24] an open-access image database, and Google images. Obtained a dataset of about 5,100 images and divided into class five: diseases caused by Alternaria Solani as in Fig. 4. healthy as in Fig. 5. Phytophthora Infestans as in Fig. 6. Viruses as in Fig. 7. and Insects as in Fig. 8.

Fig. 4. Alternaria Solani. Symptoms in this disease are leaves with brown, non-glossy dead spots, concentric rings in a target-board pattern.

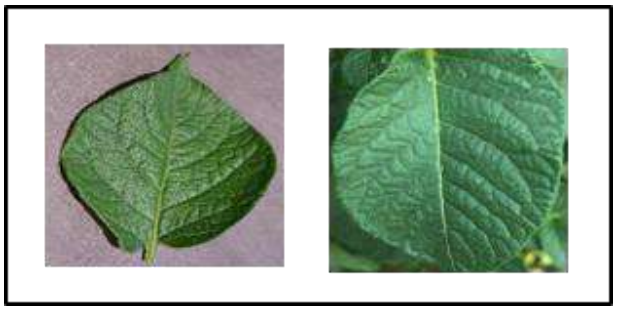


Fig. 5. Healthy.

Fig. 6. Insect. The symptom of an attack from insects is visible holes in the leaves.

Fig. 7. Virus. There are several types of viruses that attack potato leaves but the symptoms are limited by leaves that have necrotic spots.

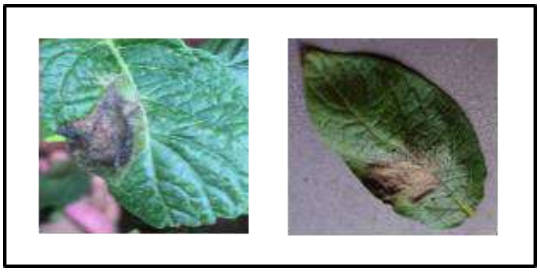


Fig. 8. Phytophthora Infestans. Symptoms of this disease are leaves with irregular, dark brown to black spots on leaves.

B. Data Pre-Processing

Pre-processing data aims to improve the quality of data to realize an accurate training model output. The first step is to minimize the noise in the image, and if there is excessive noise in the image, then it will not be used. Acquired images have a variety of sizes, and images are resized to 800x600 pixels to standardize the input of images in datasets

C. Data Augmentation

Deep Learning (deep network) requires much data when compared to the shallow network of machine learning. The lack of training data and the balance of the amount of data in each class are common problems in Machine Learning [25]. The method used to overcome this problem is data augmentation. Data augmentation is a technique of manipulating data without losing the essence of the data. Data augmentation needs to be applied in this study because 5100 datasets are still inadequate to get optimal performance. Number of augmented data that we generate is 9000 images. Many methods are used in data augmentation, ranging from simple image transformation methods such as turning, rotating, enlarging, and cropping images to the histogram-based method. In this study, only simple transformation methods such as rotating and cropping were applied.

D. Image Classification

Deep learning (DL), similarly known as deep neural learning or deep neural network, is part of machine learning (ML) in artificial intelligence (AI). The term "deep" means that Deep Learning has more layers than Machine Learning. Deep learning methods have improved the state-of-the-art in image classification, speech recognition, visual object recognition, object detection, and many other domains [7]. In Deep Learning, Convolutional Neural Network is one of the popular classes. Some studies use the convolutional neural network method to detect diseases in plants based on leaf conditions [4] [17] [18]. Convolutional Neural Networks generally consist of one or more convolutional layers that are grouped by function. Often the subsampling layer is followed by one or more layers that are fully connected as a standard neural network. Each feature layer receives input from a feature set located in a small area on the previous layer. LeNet was the beginning of the CNN architecture model and has now evolved into modern CNN architectural models such as AlexNet, VGG network, GoogLeNet, residual networks (ResNet), densely connected networks (DenseNet) [26].

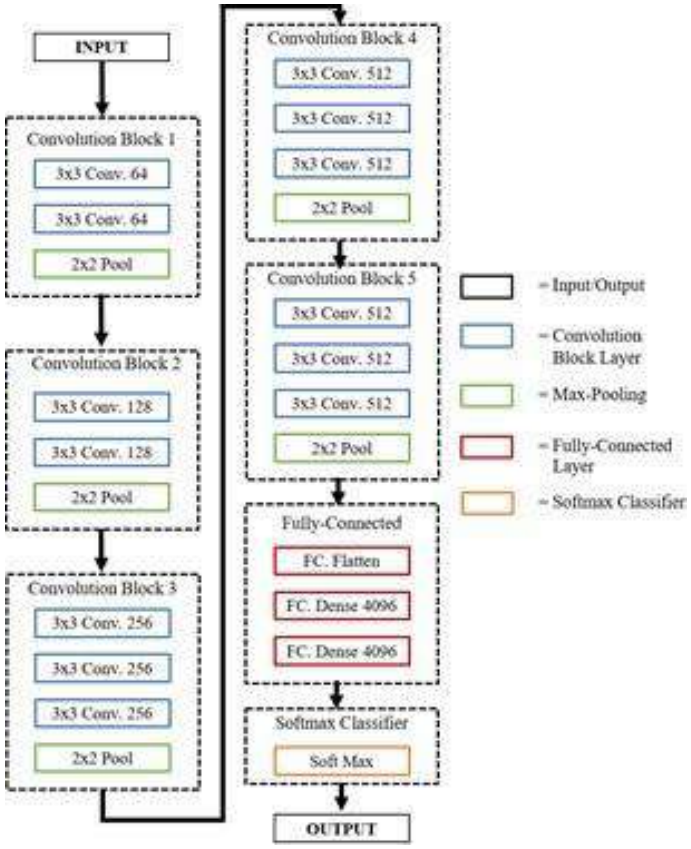


Fig. 9. VGG16 Architectural Model.

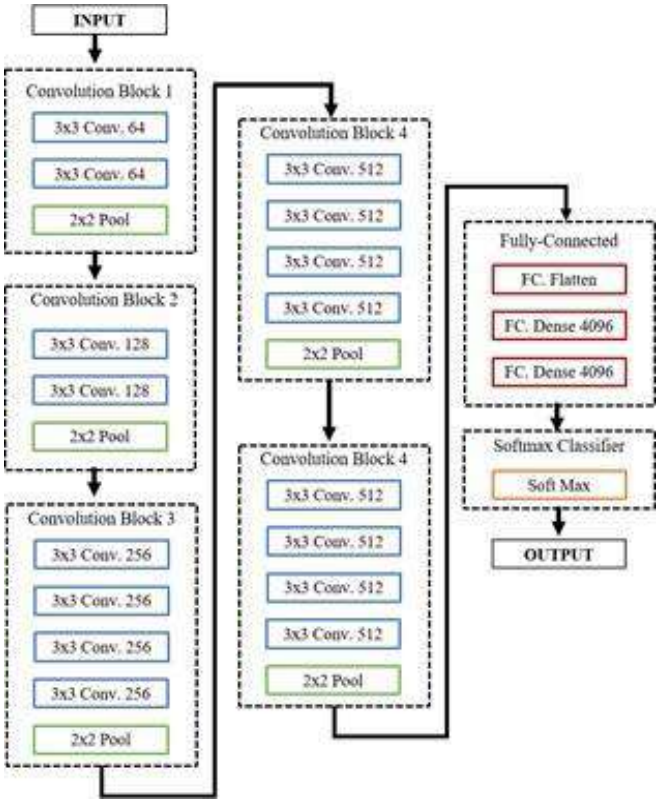


Fig. 10. VGG19 Architectural Model.

The architectural model used in this study to detect potato leaf disease is VGGNet. This architecture is the development of the AlexNet architecture model by changing the kernel size to smaller, several filters measuring 3X3 one by one. Stacked small kernels are considered more efficient and cheaper compared to large kernels. In VGGNet, the input as in Fig. 1. Passed through 5 convolutional layer blocks where each block consists of an increase in the number of 3x3 filters. The ReLU activation layer is applied in each block to recognize non-linearity so that the model can more easily adapt to various data. Dropout management techniques are used with a fixed probability of 0.25 to reduce overfitting conditions. Max-pooling layer separates blocks. Three fully connected (FC) layers follow five convolutional layer blocks. The softmax layer is the last layer to produce class probabilities. VGGNet evaluates very deep convolution networks in up to 19 layers. In this study, the authors used the two most popular types of VGGNet. VGG16 architecture shown in Fig. 9. contained 16 layers while VGG19 architecture shown in Fig. 10. contained 19 layers. The difference between the VGG16 architectures and VGG19 architecture is in the number of layers, as shown in Fig. 7. and Fig. 8. The proposed method

seeks to come across values in the image dataset to be able to recognize new images. Each step of the epoch must be the same as the value of the image being trained. The results of the epochs will be recorded to determine the value of loss and accuracy. Loss is an indication of a bad value on the model, the value of the loss obtained must be close to zero or equal to zero and accuracy value is a parameter of the success level of the system in classifying objects.

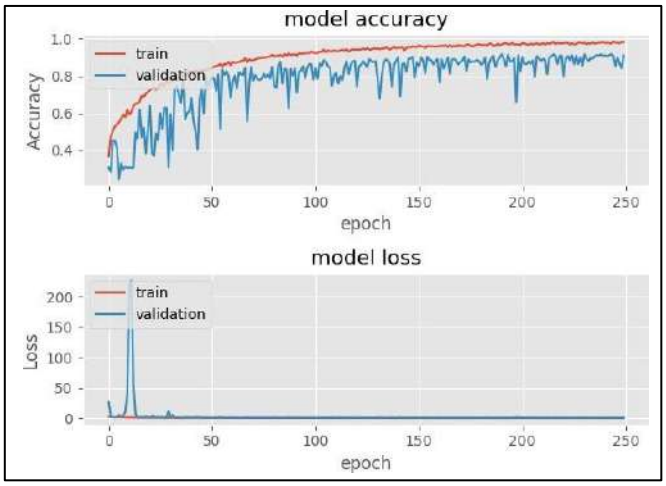


Fig. 11. The Plot of Accuracy and Loss using VGG19.

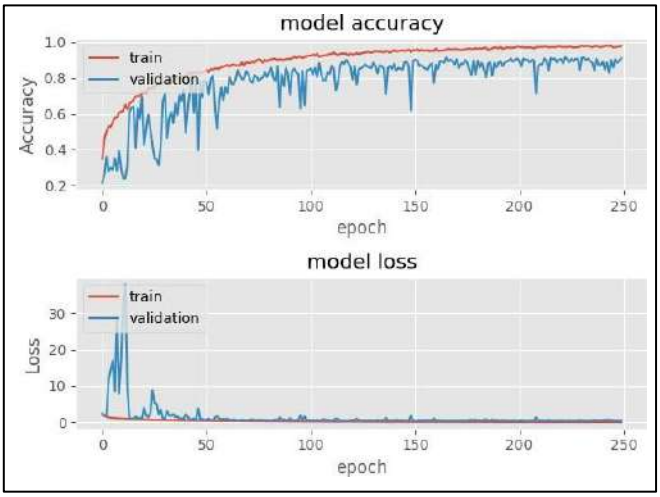


Fig. 12. The Plot of Accuracy and Loss using VGG16.

The results of the Loss values are presented in both plots using VGG19 in Fig. 11. and VGG16 in Fig. 12. getting values close to zero. VGG16 had more oscillation values in the early epoch.

The accuracy value presented in both architectures related to training with all datasets containing original images after the 250th epoch reaches 91%.

EXPERIMENT RESULT

A. Training Process

The dataset learning experiment was carried out using the Neural Network Convolutional method using the VGGNet family architecture model, specifically VGG19, which has 19 layers and VGG16, which has 16 layers. The epoch specified in this study was 250 epochs with 64 batch size and a learning rate is 0.01 to improve the performance of the model. The learning process means that the algorithm with Data augmentation is added in this study to enrich variations of the potato leaf dataset used with the same architecture. In this part, using the VGG16 architecture with a dataset that developed from collected datasets before, augmentation gives a graph of loss and accuracy result that different, shown in Fig.13. The accuracy obtained by adding augmentation data is about 93%.

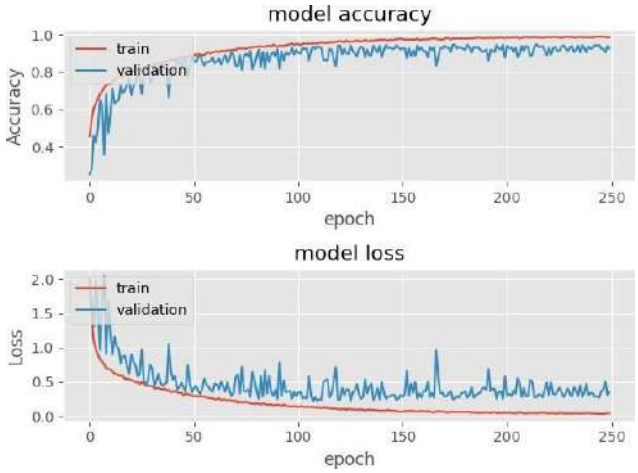


Fig. 13. The Plot of Loss and Accuracy using Data Augmentation.

The results of the system that has been done will be validated to measure the performance of a classification system that provides performance information from the resulting model. Some terms represent the results of the classification process, including True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Based on TP, TN, FP, and FN values. Then the accuracy, precision, F1 score and recall value can be obtained. The validation performed through all the validation datasets and the result of precision and recall.

B. Testing Process

After conducting the training process on the collected datasets, we retrieve the new data outside the training datasets to be tested with each model; VGG16, VGG16 with data augmentation and VGG19.

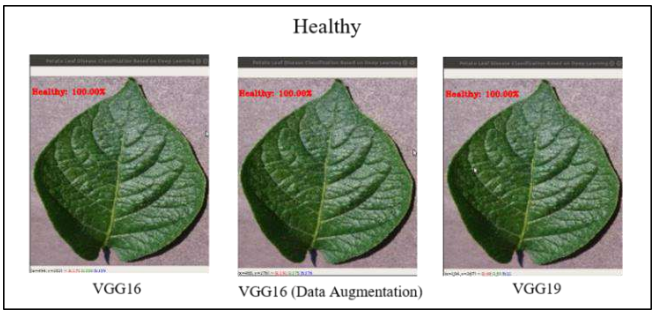


Fig. 14. Result of Healthy Leaf Testing.

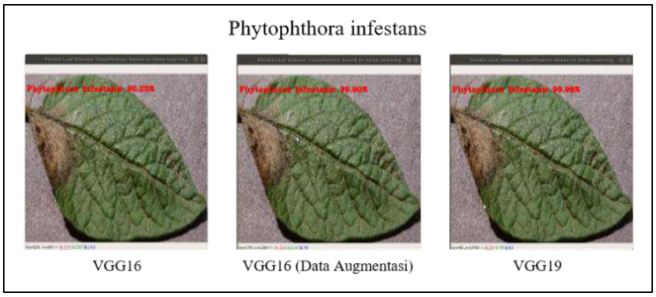


Fig. 15. Result of Phytophthora Infestans Testing.

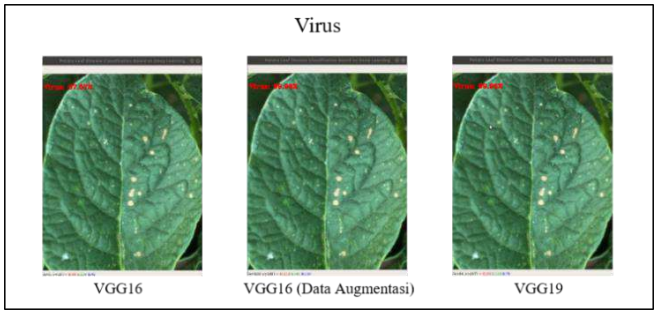


Fig. 16. Result of Virus Testing.

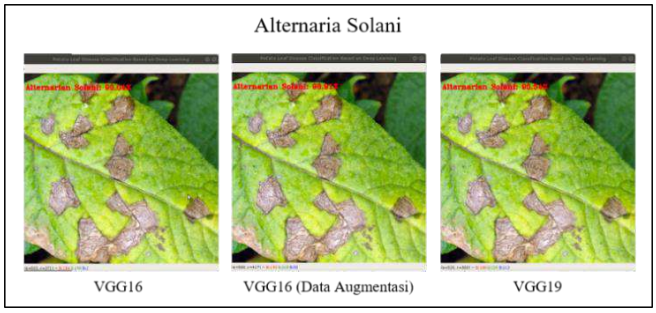


Fig. 17. Result of Alternaria Solani Testing.

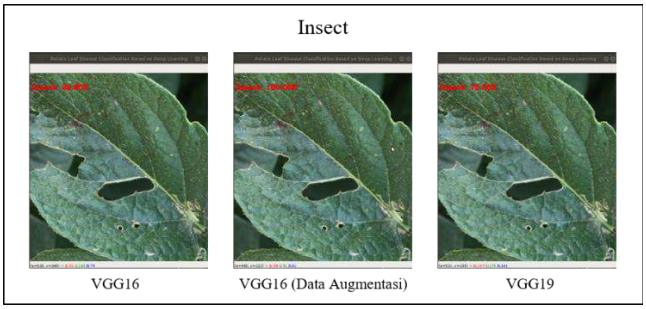


Fig. 18. Result of Insect Testing.

In this experiment, the effect of data augmentation was excellent. Most of the accuracy of the testing results with the VGG16 model using data augmentation is higher than testing using the VGG16 model without data augmentation. This is because data augmentation can multiply datasets with diverse variations without reducing the essence of the image. Besides, the difference in the accuracy of testing results from VGG16 and VGG19 is not significant. Whereas, in VGG19 layers, it has three more layers than VGG16 that make the training process using VGG19 spend more time than using VGG16. Some data shows that using VGG16 with data that is segmented produces better predictions.

It is imperative to note that when testing the model is the precision of the leaf image object, noise in the image will result in poor classification results. Minimization of noise in an image can be done by cutting off parts of the image other than the leaf object to be identified and making sure there is only one leaf in one frame.

CONCLUSION AND FUTURE WORK

The world demand for potatoes has increased significantly, mainly due to the world pandemic coronavirus. In this paper, we have presented the classification of leaf diseases from potato plants. VGGNetwork (VGG16 and VGG19) appears to be potential for studying effective features for image classification of leaf diseases. Added the data augmentation process in the dataset would produce a more robust system. Experiments show that our proposed method can achieve an average accuracy of 91-93%. We believe this work can bring many benefits in agriculture- related to world food security.

Digitalization increasing across all the fields and it is high time to adopt digitalization into the field of agriculture as well to obtain better protection in terms of growth and yield. Keeping this intention as the motivation for the proposed model to detect and classify the affected and unaffected leaves of potato. The proposed framework able to achieve an accuracy of 95.99%. Yet, this accuracy needs to be improved. The existing work further can be extended by using artificial neural networks, particularly, convolutional neural networks. These days, a lot of research related to images is happening based on CNN methodologies to obtain better and reliable accuracy. The concept of activation functions, batch normalizations, convolutional layers, and fully connected layers are playing a key role in CNN architectures to attain better accuracy.

REFERENCES

1. Statistics Indonesia, 2018 Inter-Census Agriculture Survey (SUTAS), Jakarta: Statistics Indonesia, 2018.
2. Ministry of Agriculture's Food Security Agency, 2017 Annual Report of the Ministry of Agriculture's Food Security Agency, Jakarta: Ministry of Agriculture's Food Security Agency, 2018.
3. K. . A. Beals, "Potatoes, Nutrition and Health," American Journal of Potato Research, no. 96, pp. 102-110, 2019.
4. P. TM, P. Alla, K. S. Ashirta, N. B. Chittaragi and S. G. Koolagudi, "Tomato Leaf Disease Detection using Convolutional Neural Network," International Conference on Contemporary Computing (IC3), 2018.

5. S. Sladojevic, M. Arsenovic, A. Andras, D. Culibrk and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Computational Intelligence and Neuroscience, p. 11, 2016.
6. K. Andreas and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," 2018.
7. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521 no. 7553, p. 436, 2015.
8. A. Risnumawan, I. A. Sulistijono and J. Abawajy, "Text detection in low resolution scene images using convolutional neural network," in International Conference on Soft Computing and Data Mining, Bandung, 2016.
9. M. L. Afakh, A. Risnumawan, M. E. Anggraeni, M. N. Tamara and E.
10. S. Ningrum, "Aksara jawa text detection in scene images using convolutional neural network," in 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES- KCIC), 2017.
11. I. A. Sulistijono and A. Risnumawan, "From concrete to abstract: Multilayer neural networks for disaster victims detection," in 2016 International Electronics Symposium (IES), 2016.
12. I. A. Sulistijono, T. Imansyah, M. Muhajir, E. Sutoyo, M. K. Anwar,
13. E. Satriyanto, A. Basuki and A. Risnumawan, "Implementation of Victims Detection Framework on Post Disaster Scenario," in 2018 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2018.
14. M. K. Anwar, M. Muhajir, E. Sutoyo, M. L. Afakh, A. Risnumawan,
15. S. Purnomo, E. S. Ningrum, Z. Darojah, A. Darmawan and M. N. Tamara, "Deep Features Representation for Automatic Targeting System of Gun Turret," in 2018 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2018.
16. H. Imaduddin, M. K. Anwar, M. I. Perdana, I. A. Sulistijono and A. Risnumawan, "Indonesian Vehicle License Plate Number Detection Using Deep Convolutional Neural Network," in 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), 2018.
17. D. M. Dinama, Q. A'yun, A. D. Syahroni, I. A. Sulistijono and A. Risnumawan, "Human

Detection and Tracking on Surveillance Video Footage Using Convolutional Neural Networks," in 2019 International Electronics Symposium (IES), 2019.

18. A. Risnumawan, M. I. Perdana, A. H. Hidayatulloh, A. K. Rizal, I. A. Sulistijono, A. Basuki and R. Febrianto, "Automatic Detection of Wrecked Airplanes from UAV Images," International Journal of Engineering Technology (EMITTER), vol. Vol 7 no 2, pp. 570-585, 2019.
19. S. Sladojevic, M. Arsenovic, A. Andras, D. Culibrk and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Computational Intelligence and Neuroscience, 2016.
20. E. Fujita, Y. Kawasaki, H. Uga, S. Kagiwada and H. Iyatomi, "Basic investigation on a robust and practical plant diagnostic system," 2016 15th IEEE International Conference on Machine Learning and Applications, pp. 989-992, 2016.
21. P. Raja, C. Szczepanski, K. V. Mukesh, R. Ashiwin and R. Aniirudh, "Disease Classification in Maize Crop using Bag of Features and Multiclass Support Vector," Second International Conference on Inventive Systems and Control, pp. 1191-1196, 2018.
22. P. Padol and A. A. Yadav, "SVM Classifier Based Grape Leaf Disease Detection," 2016 Conference on Advances in Signal Processing (CASP), pp. 175-179, 2016.
23. J. Francis, A. S. Dhas D and A. B. K, "Identification of Leaf Disease in Pepper Plant Using Soft Computing Techniques," Conference on Emerging Devices and Smart Systems (ICEDSS), 2016.
24. E. Hosaain, M. F. Hossain and M. A. Rahaman, "A Color and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier," International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.
25. A. Rastogi, R. Arora and S. Sharma, "Leaf Disease Detection and Grading using Computer," 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015.
26. S. P. Mohanty, D. Hughes and M. Salathe, "Using Deep Learning for Image-Based Plant," 2016.
27. A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning,"

International Interdisciplinary PhD Workshop (IIPhDW), pp. 117-122, 2018.

28. A. Zhang, A. J. Smola, M. Li and Z. C. Lipton, Dive into Deep Learning, 2020.

PAATHSHALA: A VIRTUAL CLASSROOM ON THE PHASE OF PREVALENT EPIDEMICS IN THE CHANGING WORLD

Shivangi Chauhan[1], Shubham Gola[2], Kush Sharma[3], Gaurav Gahlawat[4] , Gaurav Dubey[5]

Deparment of Computer Science Engineering,

Mangalmay Institute of Engineering And Technology, Greater Noida, UP, India

ABSTRACT

In the last many decades, education has witnessed some advances in technologies involving computer-backed literacy that promises to drastically change the styles of tutoring and literacy. The World Wide Web has played a major part in information storehouse and dispersion in the educational community. Conventional classroom-grounded tutoring involves the delivery of course accouterments by the speaker in a particular place at a defined time. Hence it imposes a constraint of time and place on both the educator and the pupil. Due to mortal factors arising from the traditional classroom system, the speaker may not always be suitable to put in optimum trouble towards preparing and delivering course accouterments. There may also be inconsistencies in the pedagogy and literacy style due to the repetitious nature of tutoring/literacy. The idea of this paper is to develop a virtual classroom system to enhance learning on the lot. The system was developed using PHP and MySQL as garçon side programming and database independently. The web-grounded virtual classroom provides a web-enabled interactive model fore-learning in which the course material is presented using multimedia and hypermedia.

Keywords: *Virtual classroom, e-learning, multimedia, education.*

INTRODUCTION

With the added use of network computers, the Internet, and advances in telecommunication technology, e-Learning has been extensively honored as a precious tool for literacy and training. The traditional means of advanced education have remained dominant in seminaries in some developing countries. With the significant growth of learning, preceptors and scholars typically explore new ways of constructing knowledge. The current technology being heavily delved into as an educational platform is the World Wide Web(WWW). The WWW which represents a

platform for information storehouse and dispersion can be penetrated in minimal time, and this is veritably important to the educational community. The fact is that the transition from a digital peak society to a global vill information society causes the traditional educational model to be unfit to cover the educational requirements of ultramodern societies. The globe is faced with a transition from a static frugality to a new knowledge-driven frugality.

Population explosion and adding admission requests into seminaries in every region of the world brought lesser constraints on the coffers of several seminaries. In case, there's a problem of a shy number of mortal and material coffers to feed the education of the large population. The population of academy-age citizens in utmost places has grown extensively to the extent that only a small chance can be offered admission. A new literacy terrain needs to be created which will give autonomy and inflexibility, establish connections and easy communication between centers of culture and knowledge, and grease easy access for all citizens of a knowledge-grounded society. Conventional classroom-grounded tutoring involves the delivery of course lectures by the speaker in a particular place at a specific time. Hence it imposes a constraint of time and place on both the educator and the pupil. Due to the mortal factor, the speaker may not always be suitable to put the optimum trouble towards preparing and delivering course models.

The remedy to this situation seems to be the literacy ways that are grounded on ultramodern technologies similar to the Internet and WWW combined with traditional classroom tutoring. One of the ways this can be achieved is through the use of virtual classrooms. A virtual classroom is a terrain conducive to literacy, which takes place in cyberspace. It provides the tools that learners need and brings together preceptors and learners to partake in information and ideas. A virtual classroom is a special form of relearning that finds applicable operations in perfecting the conventional literacy styles editorialized that learning can be stationed using a wide range of technologies and media.

The virtual classroom has its roots in the study of computers in education similar to computer-intermediate instruction and multimedia as an educational tool. These broad fields covered not only hypermedia, similar to web-grounded hypertext but also on-internet educational software design ranging from media academy surgery tutorials to interactive CD-ROM terrain atlases. Numerous of the issues facing these virtual classrooms, similar to the evaluation of interface

design, integration of computers into course design, and social issues of computing are largely applicable to the design and use of internet-grounded virtual classrooms.

Present technologies enable the creation of virtual classrooms using the Internet and its coffers. For the preceptors and trainees, a benefit of the Internet as a platform for the virtual classroom is that the information that can be stored is nearly measurable. One of the benefactions of a Virtual Classroom(VCR) is that access to high-quality and flexible literacy technologies. The information being electronically stored can be penetrated or downloaded by learners at their own pace, thereby booting the constraint of time and place endured in classroom- grounded literacy. The involvement of distance literacy includes tutoring using telecommunication tools, which transmit and admit multitudinous accouterments through data, voice, and videotape. There's also an increased use of virtual classrooms(online donations delivered live) as an online literacy platform and classroom for a different set of education providers. In addition to virtual classroom surroundings, social networks have come an important part of relearning.

LITERATURE REVIEW

Quite a lot of studies live relating toe-learning, distance literacy, and virtual literacy. These terms are occasionally used interchangeably. According to(5), e- learning means literacy that makes use of a network for delivery, commerce, or facilitation. This type of literacy includes distributed learning and distance literacy. Computer- Grounded Training(CBT) is delivered over a computer network and web-grounded training(WBT). It may be computer-grounded, coetaneous, asynchronous, educator- grounded or a combination of the forenamed.

Former workshops in the area of the virtual classroom will be bandied in this section following their literal development of VCR, architectural design and system perpetration, and provision of e-learning platforms for the impaired. The paper in(4) addresses the history of distance literacy, current issues, the civil government’s part, and four specific areas of enhancement including classes change, new patterns of commerce, changes in organizational structures, and the places and conditioning of actors in both business and academic distance- learning surroundings.

A model for perfecting online educational systems for both preceptors and learners was proposed in (11). The model allows for more accurate assessment and further effective evaluation of the literacy process. The model includes logistics systems to show that it could be necessary to integrate systems that handle a payload of handbooks and other physical accouterments to

distance scholars. The study in (6) discusses the architectural design of an intertwined system for the delivery of lectures in a virtual terrain. The armature and description of the system factors are presented with the ways and recommendations for the perpetration of the designed system. The system armature is multi-tier, modular, scalable, and erected for rigidity to the database middleware suite. All functionalities within the operation are delivered using web services, communicate via assiduity standard XML messaging and access is pure via a web cybersurfer. The study in(7) discusses gests in developing VCM with different authoring tools and evaluates their effectiveness. The results of the check shows that this exploration proved that the replier scholars veritably well entered the Virtual Classroom Module(VCM) developed.

Related Works 1. Piazza

Piazza is a literacy operation system that allows scholars to ask questions in a forum type format. preceptors are suitable to moderate the discussion, along with championing accurate answers. The software was constructed by Pooja Nath in 2009 in order to speed response times and produce a common place where scholars could engage in discussion outside of the classroom. Utilising an expansive announcement system and a simple layout, the response time on Piazza pars roughly 14 twinkles. preceptors also have the capability to allow scholars to post anonymously, encouraging further in-depth discussion. druggies can intimately(and anonymously, if the head educator allows it) ask questions, answer questions, and post notes. Each question prompts a collaborative answer to which any stoner can contribute and an educator answer, shown directly below, which can only be edited by preceptors. Multiple scholars are allowed to contribute to each answer like Wikipedia entries, and each answer has a interpretation history that shows what each pupil wrote. druggies are allowed to attach external lines to posts, use LaTeX formatting, view a post's edit history, add followup questions, and admit dispatch announcements when new content is added. The interface consists of a dynamic list of posts on the left side of the screen, a central panel for viewing and contributing to individual posts, and an upper bar for account control. According to the company's data, the average Piazza question is answered within 14 twinkles. Individual Piazza classes are tone-contained and can be locked with an access law. Anyone may produce a class, but the head educator retains full control over the class content, along with executive capacities similar as championing good answers and viewing more detailed statistics on class exertion.

2. Google Classroom

Google Classroom is a free amalgamated literacy platform developed by Google for educational institutions that aim to simplify creating, distributing, and grading assignments. The primary purpose of Google Classroom is to streamline the process of participating lines between preceptors and scholars. As of 2021, roughly 150 million druggies use Google Classroom. Google Classroom integrates a variety of other Google Applications for Education, similar to Google Docs, Google wastes, Google Slides, Gmail, and Google timetable into a cohesive platform to manage pupil and schoolteacher communication. scholars can be invited to join a class through a private" class law", or be imported automatically from an academy sphere. preceptors can produce, distribute and mark assignments all within the Google sphere. Each class creates a separate brochure in the separate stoner's Google Drive, where the pupil can submit work to be graded by a schoolteacher. Assignments and due dates are added to Google timetable, where each assignment can belong to an order or content. preceptors can cover each pupil's progress by reviewing the modification history of a document, and after being graded, preceptors can return work along with commentary and grades.

Proposed Methodology

Some of the initial features of the application will be as follows:

- Simple, intuitive, and interactive UI
- Easy Onboarding
- A personalized Classroom web application that keeps data secure by not using any third-party application.
- Supports multi-user tier of admin, faculty, class representative, and students.
- In-house assignment submissions and grading.
- Dynamic real-time timetable avoiding conflict between classes, streamlining student experience.
- Easy access/download to study resources.
- Dedicated forms section for feedback, doubts, etc. with the added functionality of reminders.

- Tracking attendance of a student using an interactive chart.
- Dedicated section for placement-related information, enabling students to apply for ongoing placement drives.
- Dedicated Class space and institution space for discussions.

Motivation

The effect of an epidemic on the education sector has redounded in the shift of education mode to online. From classes to tests, everything was forced to be nearly. ignorance and randomness to the situation could affect the studies. Hence, there's a need for an operation that enables preceptors as well as scholars to maintain the class in a systematized manner. Also, it must total all the academic conditioning while keeping data secure. Conventional classroom-grounded tutoring involves the delivery of course lectures by the speaker in a particular place at a specific time. Hence it imposes a constraint of time and place on both the educator and the pupil. Due to the mortal factor, the speaker may not always be suitable to put the optimum trouble towards preparing and delivering course models.

The remedy to this situation seems to be the literacy ways that are grounded on ultramodern technologies similar to the Internet and WWW combined with traditional classroom tutoring. One of the ways this can be achieved is through the use of virtual classrooms. A virtual classroom is an terrain conducive to literacy, which takes place in cyberspace. It provides the tools that learners need and brings together preceptors and learners to partake information and ideas. A virtual classroom is a special form of learning that finds applicable operations in perfecting the conventional literacy styles editorialized that learning can be stationed using a wide range of technologies and media.

Present technologies enable the creation of virtual classrooms using the Internet and its coffers. For the preceptors and trainees, a benefit of the Internet as a platform for virtual classrooms is that the information that can be stored is nearly measureless. One of the benefactions of a Virtual Classroom(VCR) is access to high-quality and flexible literacy technologies.

The information being electronically stored can be penetrated or downloaded by learners at their own pace, thereby booting the constraint of time and place endured in classroom-grounded literacy. The involvement of distance literacy includes tutoring using telecommunication tools,

which transmit and admit multitudinous accouterments through data, voice, and videotape. There is also an increased use of virtual classrooms(online donations delivered live) as an online literacy platform and classroom for a different set of education providers. In addition to virtual classroom surroundings, social networks have come an important part of learning.

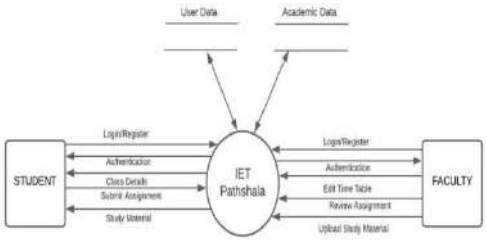
Also, an In-house Virtual Classroom brings several benefits of its own. It enhances the scalability and cost-effectiveness of the overall result. It offers customizations, which were not possible before with third-party results.

Some of the customizations are described below.

1. Devoted Institute space
2. Added up academic conditioning
3. Pupil document upload and verification
4. Virtual examinations
5. Control over institute data
6. Instant bug fix

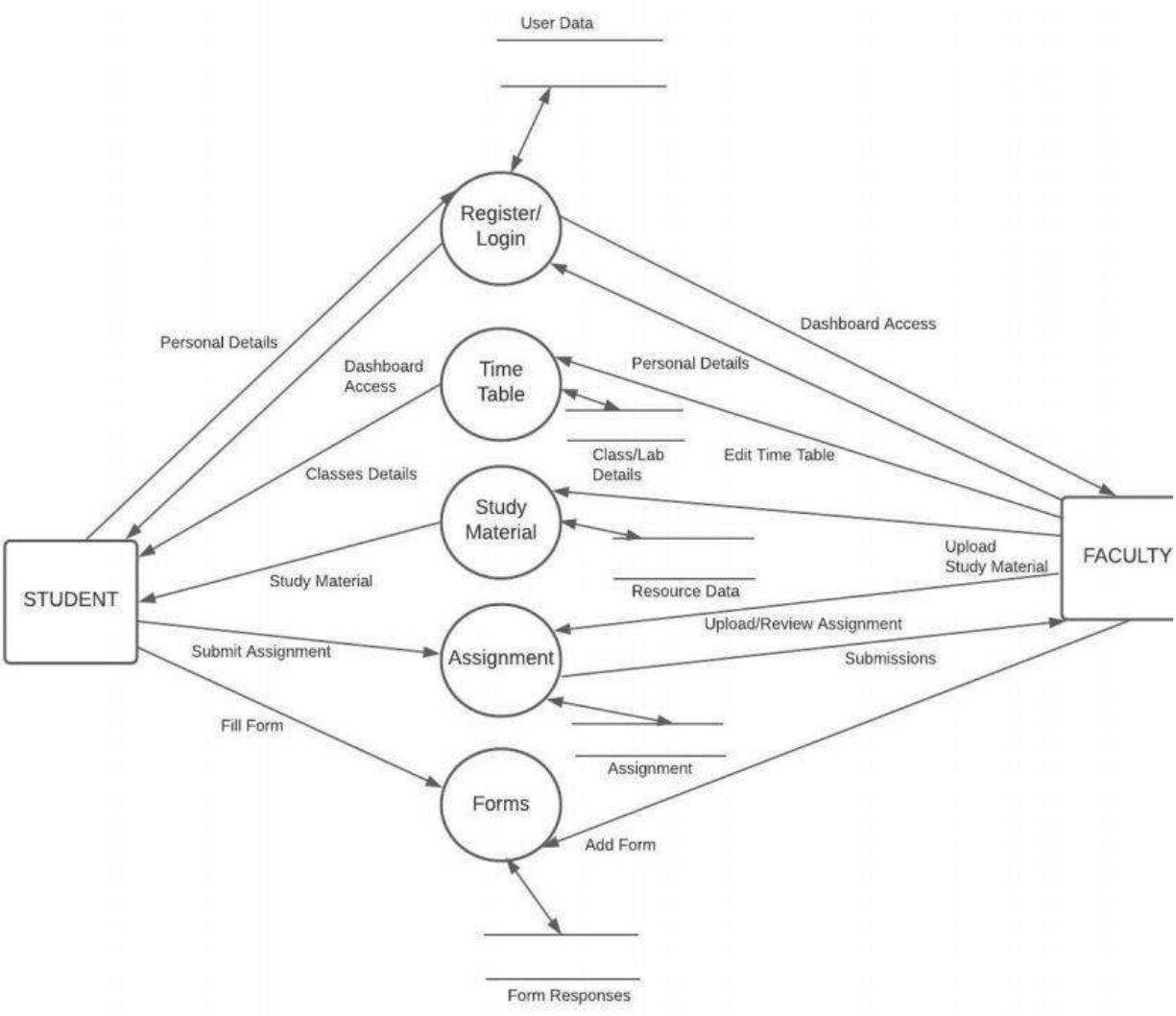
Also, It resolves the issue of data security.

PLAN OF WORK



(System Design) Data Flow Diagrams

The following two diagrams are the data flow diagrams (level 0 and level 1) of the system with two external entities: Student and Faculty.



The level-1 diagram has four processes:

Time Table, Assignment, Forms, Study Material, and data flow is being shown.

Use Case Diagram

In the following figure, the use case diagram is showing two actors: student and faculty. Students can view the timetable, submit the assignment, view study material, view forms. Faculty can add and view forms, study material, view and edit time-table, and can review assignments. References will be used for connecting between different Database models instead of encapsulating data in the student model. Some of the Database models will be as follows: -

- Assignment
- Attendance

●Chapter

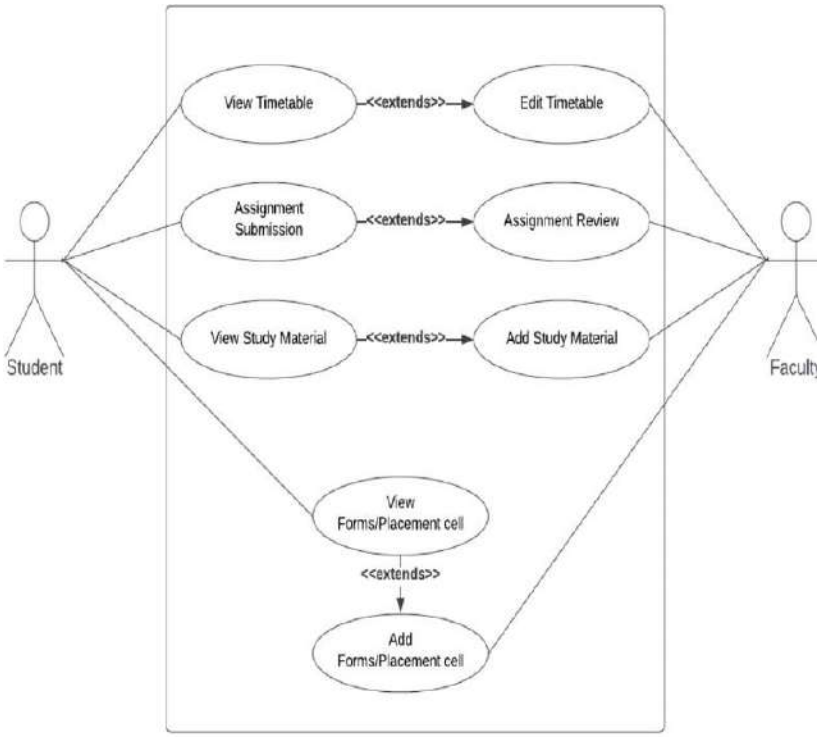
●Form

●Notification

●Reminder

●Subject

●Submission



Major technologies used in the whole process are described as follows: -

- Backend- NodeJS, ExpressJS
- GUI Frontend- HTML/CSS, Bootstrap, JavaScript
- IDE- Visual Studio Code
- Database- MongoDB Atlas
- API Usage- Google Open Source APIs
- Version Control System- Github

- Deployment- Heroku

3.3 Languages Used

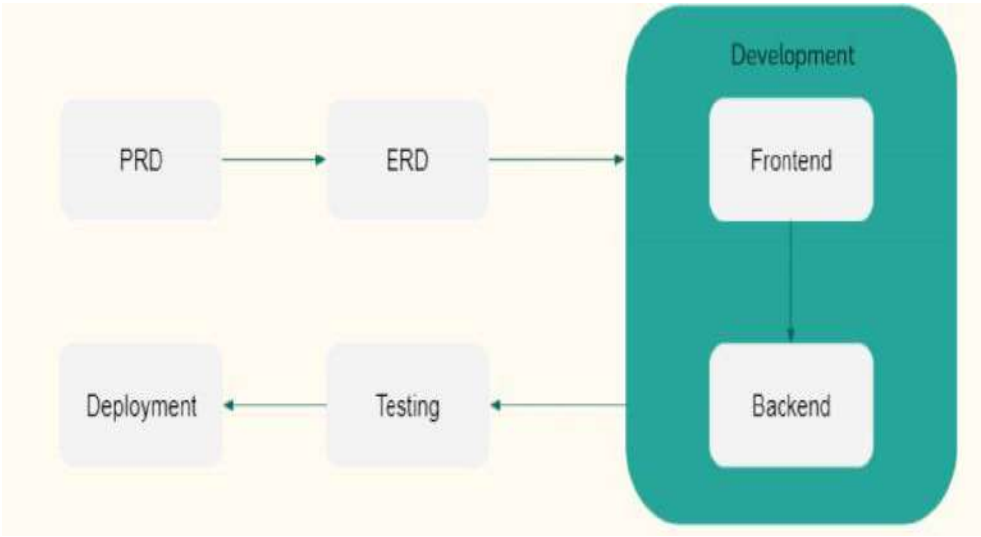


Figure 3.3 shows different stages of the Software Cycle Action plan.

- PRD (Product demand Document)

A product conditions document (PRD) is a document containing all the conditions to a certain product. It's written to allow people to understand what a product should do. A PRD should, still, generally avoid anticipating or defining how the product will do it in order to latterly allow interface contrivers and masterminds to use their moxie to give the optimal result to the conditions

- ERD (Engineering Requirement Document)

An engineering conditions document (ERD) is a statement describing the thing and purpose of a new element. Unlike a product conditions document (PRD), which tells masterminds what they need to make, an ERD specifies why a part is being erected and how its design energies its purpose. By following the engineering conditions outlined in an ERD, masterminds can insure that the part they make will satisfy client requirements.

Software Development Life Cycle (SDLC) model: -

- Nimble model /Agile Model

Nimble software development refers to a group of software development methodologies grounded on iterative development, where conditions and results evolve through collaboration between tone- organising cross-functional brigades. nimble styles or nimble processes generally promote a disciplined design operation process that encourages frequent examination and adaption, a leadership gospel that encourages cooperation, tone- organization, and responsibility, a set of engineering stylish practices intended to allow for rapid-fire delivery of high-quality software, and a business approach that aligns development with client needs and company pretensions. nimble development refers to any development process that's aligned with the generalities of the Agile Manifesto. Deployment Phases

- Product (on AWS)

The Sprint will be of duration of one month. The first two sprints will be utilized for the PRD, ERD, and UI development as it'll be a static part and makes the testing and integration process a lot easier. For the backend features, an incremental aspect of the nimble model will be used. In the prototype, some of the abecedarian features i.e. Authentication, stoner model, etc will be enforced.

In the posterior prototypes of the operation, features will be added and integrated into the result. Backend will take three devoted sprints after the first two sprints allocated for PRD, ERD, and UI development.

RESULT AND DISCUSION

In this paper, a virtual literacy system has been developed. The new system is anticipated to serve as a remedy for the problems and weaknesses observed in the old system. It'll combine open literacy ways grounded on new technologies(in this case, the world wide web) with conventional classroom tutoring, The main intention is to make the literacy experience more flexible, stimulating, and available around the timepiece and at any place with Internet installations. The scholars will be suitable to navigate freely within the virtual classroom terrain and enhance the information coffers used by the scholars.

ONBOARDING FLOW

To get a user onboard, they must register on the page (shown above) first. During the registration/sign-up, we are storing the user's information as Name, Branch, Semester, Email id, and password (to be generated), and a file of a photograph.

This will allow users to get registered with IET Path Shala. To get started, a user is required to sign-in

FORMS

Forms section is one channel with which Teachers, CRs etc. can provide some information and can take inputs from students. Example: - Feedback Form, Doubt Clearing Form, Consensus Form etc. In these forms, teachers will post the information and the link leading to the google form. Students will, then, fill their inputs in the form. This will help in taking feedback on the teaching style, to clear doubts, and to take students' consensus on any topic in an organized manner.

CONCLUSION AND FUTURE WORKS

IET Pathshala is very useful in many aspects for the students as well as teachers of the institutions to aggregate all the activities at one platform in an organised manner. Through this, we are providing the facility of scheduling online live classes through google meet and providing study material, assignment posting and grading etc. Though this is currently solving several issues but can be improved in many dimensions. There will be a messaging application in the classroom where students can form different communities as per their interests and connect with each other, they can share their views and opinions, future career opportunities etc.We can also include a section for contests and competitions organised byteachers/seniors/alumni.

The students can take part in these competitions/hackathons and getinspired to perform better.An open channel, in which students can post their doubts, queries etc. and their fellow mates can solve them or help them, will be inculcated in the application. This will help to increase the spirit of self help and mutual help among students. A section of E-Library should also be there where students can access e- books of the curriculum as well as other books to help students to grow in a holistic way.It will be a step to save paper as well.

As of now, we are using google meet for the live classes, meetings etc. In the future, we can also try to build a video conferencing application of our own. It will also reduce the issues of data security and make our application self reliant. Due to the constraint of time as well as resources, we could not implement these features but this part could be done in the future as it will widen the scope of the application.

REFRENCES

1. Dr N. K. Jain (VCR Coordinator), Virtual Classroom (VCR), IIT Indore. Project implementation for The use of Virtual Classroom technology alongside the college's Virtual Learning Environment (VLE).
2. Dilani S. P. Gedera, Students' experiences of learning in a virtual classroom, International Journal of Education and Development using Information and Communication Technology. (IJEDICT), 2014, Vol. 10, Issue 4, pp. 93-101. 1180
3. Akinyokun Oluwole Charles, Iwasokun Gabriel Babatunde., "Design and Implementation of a Web-Based Virtual Classroom System". IOSR Journal of Research & Method in Education (IOSR-JRME) e-ISSN: 2320–7388, p-ISSN: 2320–737X Volume 4, Issue 3 Ver. II (May-Jun. 2014), PP 68-77.
4. Kimberly C. Harper, Kuanchin Chen, David C. Yen. Distance learning, virtual classrooms, and teaching pedagogy in the Internet environment. Elsevier. Technology in Society 26 (2004) 585– 598.
5. Cervino, J. (2007), The Virtual Classroom available online at <>.
6. A.I. Obasa, 2A.A. Eludire and 1Mbing Isaac. The Architectural Design of an Integrated Virtual Classroom System Research Journal of Information Technology 3(1): 43-48, 2011 ISSN: 2041- 3114. © Maxwell Scientific Organization, 2011 Pp 43-48.
7. Dr. Nayereh Shahmohammadi. Learning with Virtual Classroom Module, how Effective?. Research and Educational Planning Organization, Ministry of Education. Journal of Applied Science and Agriculture, 8(3): 269-274, 2013 ISSN 1816-9112.
8. Birgit Rognebakke KROGSTIE Introducing a Virtual Classroom in a Master Course: Lessons Learned. The work was conducted as part of the Socrates Minerva project "Virtual Classrooms in European Provision"

(http://learning.ericsson.net/virtual/products.shtml) which aims to develop best practice-founded guidelines for the use of virtual classrooms in European organisations, public and corporate.

9. Mohammad Hassan Falakmasir, Jafar Habibi Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning. Internal Educational Data Mining Society (2010).
10. Florence Martin, Michele A. Parker Use of Synchronous Virtual Classrooms:
11. Why, Who, and How? MERLOT Journal of Online Learning and Teaching Vol. 10, No. 2, June 2014.
12. P. Nagarajan, Dr.G.Wiselin Jiji ONLINE EDUCATIONAL SYSTEM (e- learning), International Journal of u- and e- Service, Science and Technology. Vol. 3, No. 4, December, 2010.

A STUDY ON ENVIRONMENTAL PROTECTION BY ADOPTING CAR POOL SHARING TO ASSIST NATIONAL DEVELOPMENT

Pratyush Raj[1], Akhil Kumar[2], Anshul[3]

[1,2] Student, Mangalmay Institute of Engineering & Technology

[3] Assistant professor, Mangalmay Institute of Engineering & technology

ABSTRACT

This article presents the design and implementation of a ride sharing application for a mobile environment. It will enable users to share rides in an efficient and simple way. Use of this system should reduce significantly the number of private cars on the roads, providing ecological, economical, and social benefits. The system is designed for any device, thus enabling implementation of the sharing any time, from anywhere, anytime. The system requires an algorithm for finding routes in a user-defined path, according to the source and destination along the path. This system differs from the existing ride sharing in sever always.

INTRODUCTION

Population growth and increasing population density, particularly in metropolitan areas, have brought about an increase in the number of vehicles on the roads, by a few percentage points per year (3.6% increase in 2010 alone). The cumulative effect of this phenomenon is staggering. The main derivatives of this situation include (in addition to direct economic expenditures on car maintenance, insurance and fuel):

TrafficCongestion

On average, travelers in Delhi, Mumbai, Bengaluru and Asian cities during peak traffic times. India`s biggest cities may be losing up to $22 billion annually to traffic congestion, and its commutes are bearing the burden.

Parking

Parking is another obvious problem in large, crowded cities. Various solutions have been proposed, but due to increasing number of vehicles and population per year have much effect on

parking. Also, parking is becoming expansive. e.g., "fast lane", which offers free passage to vehicles with 4 or more passengers. In London, a heavy daily fee is exacted from commuter cars entering the city center.

Environmental Concerns

Congestion has heightened the awareness of the importance of environmental protection and there is a worldwide search for new, energy efficient ways to manage our daily mobility. Unfortunately, none of these efforts made a significant contribution to the situation.

LITERATURE REVIEW

The term "carsharing" can be defined as "a service that provides members with access to a fleet of vehicles on an hourly basis". The first large-scale car sharing program was implemented in Switzerland in 1987, and the first car sharing organization in the United States appeared in 1998 in Portland, Oregon. As of January 2008, 18 car sharing services were operating in the United States with a combined membership of approximately 234,834 people and over 5,200,600 cars in use (McLaughlin, 2008). Brook (2008) reported that as of March 2008, nearly100 U.S. cities have some sort of formal car sharing operation. [1]

Car Sharing is considered as a short-term car rental, allowing members to gain the benefits of private car use without the costs and responsibilities of ownership. According to Navigant Research, the worldwide number of cars sharing members will continue to grow from 2.3 million in 2013 to more than 12 million by 2020.Carsharing services revenue is estimated to grow from approximately $1 billion in2013 to $12 billion by 2020. In order to improve overall efficiency, user-friendliness and operational manage ability of car sharing service.

Car sharing has been associated with a variety of social and environmental benefits. Interms of alternatives to the private automobile, car sharing has been described as the "missing link" because it provides users greater flexibility than public transit and per-day rental cars and enables them to travel longer distances than they can byfoot, bicycles, or taxis. Car sharing also helps to mitigate environmental degradation and emissions production. First, it provides "mobility insurance" to users while they satisfy their daily travel needs via other modes; as a result, car sharing has been shown to encourage individuals to avoid purchasing new private cars or to sell cars they currently have. This helps to reduce demand for the production of new automobiles,

which consumes energy, water, and raw materials and produces hazardous emissions and waste products. Second, users have greater incentives to “trip-chain” and reduce impulsive trips because car sharing highlights the costs per car trip and requires users to plan their trips further in advance. Studies have shown that members of car sharing programs drive significantly less than non-members, which reduces the amount of carbon dioxide and other noxious emissions that enter the atmosphere. Car sharing offers numerous other benefits, including reduced parking demand, reduced traffic congestion, and opportunities for members to save money otherwise spent on car ownership costs.

PLANNING OF WORK

The system is built in Visual Studio Code, using HTML (Hypertext markup language), CSS (Cascading Style Sheet) & Bootstrap as frontend, MySQL as backend and php (Hypertext Preprocessor). It has been tested using the local server made on laptop. We chose to use above technologies for multiple reasons. First, it is supported by various types of devices. Second, they more reliable. It can run on both system and smartphones. [2]

System

In this section we describe our system in which first user have to signup and create their account and then sign in with their new account and as per there chose, they can choose between “offer a ride” and “find a ride” If they are rider than they should give two addresses as inputs, and searches for the driver to travel on the route or between those routes. This is done by iterating and examining database of those two addresses and storing all found drivers. This concept narrows down the number of drivers that are a possible meeting point between the two locations. The system works bidirectional: first, it matches the passenger’s given address to the driver’s address, and then does to opposite. If the driver accepts the request of the passenger than the chat option for both the driver and the passenger will be open and they can chat with each other about the ride and decide what to meet and at what time and if they are not satisfied by the arrangement, they see for other rides.[4]

PHP

We have used PHP in our project for sending request to server and finding response from server. PHP is very important server scripting language for our WORK. Using PHP, we communicate

Using php we send our data of our pages on data base. The following pages are sign up page find using Php we retrieve data in profile page and showing response of rider to driver of booking page after that driver accept request or reject the booking it depends on driver.

If driver accept request, then response goes back to rider. Rider and driver both can see the response of each other.[3]

There is communication network in which we also have used php for sending and showing messages that is chat box. Using chat box, they can communicate with each other. They can know each other by talking.

MySQL is the database that we use in our project with PHP. MySQL is also very important data base language for creating database add tables.

The format of storing data in database using MySQL language is very nice. For this PHP works very easily and very compatible for it.

Even now in the world use of PHP much more than other server languages because of its simplicity and compatibility with MySQL on Apache Server.

METHODOLOGY

Home Page before Sign in or Sign up

FUTURE WORK

There are some research possibilities to further work that can be added to the current application:

- Voice recognition: How convenient would that be if we could just talk to our phone

and say in our own words: "I’ m driving from Me mph is Tennessee to Chicago Illinois at 5pm, 2 available seats". Since our application is built in a modularly way, it would be very easy to add unique and special features like this. Using voice recognition will substantially increase the use of people. As technology becomes more and more advanced, we expect things to be faster and easier to handle. Talking to a phone in a free manner is, generally speaking, more user friendly, and will save him precious time.

- Account Ratings: In order to increase and attract more users the system must have a way to rate users with ratio to their use of the application [11]. The more they use the application the more benefits they should receive. Adding a system that gives points for each drive, points for good service to the hitch-hiker, will come into considerations when the hitch-hiker enters his feedback from the drive.

REFERENCES

1. https://www.research.com
2. A. M. Amey, “Real-time ridesharing: exploring the opportunities and challenges of designing a technology-based ride share trial for the MIT community.
3. https://www.guru99.com/what-is-php-first-php-program.html
4. https://www.studytonight.com/php/introduction-to-php
5. https://www.siteground.com/tutorials/php-mysql/mysql/

A PROTOTYPE ERP AND ITS BENEFITS TO INSTALL FACE RECOGNITION ATTENDANCE SYSTEM AND ITS USAGES IN THE INDUSTRY

Vinay Pratyush[1], Prashant Kumar[2], Md. Mushfique Raza[3], Saurav Kumar[4], Anshul[5]

[1,2,3,4] Student, Mangalmay Institute of Engineering & Technology, Greater Noida

[5]Assistant Professor, Mangalmay Institute of Engineering & Technology, Greater Noida

ABSTRACT

This project is focused on creating an automated attendance system for colleges and universities. With the help of CCTV cameras and facial recognition techniques. Taking attendance of students is mandatory but very time-consuming task. Important time of class is wasted on taking attendance manually. Manual attendance taking is also very erroneous process. Biometrics systems are used in many places, but in such systems long ques can form, if number of students is large. Many colleges have implemented facial recognition attendance system but often they fail in different lighting conditions and very often give false positives. Many false positives also happens because facial recognition techniques cannot differentiate between identical twins. To tackle these issues, we have made this project. We have implemented HOG (Histogram of Oriented Gradients) algorithm, developed by Robert K. McConnel of Wayland Research Inc., as feature descriptor and SVM (Support Vector Machine) algorithm, developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis, and CNN (Convolutional Neural Networks) for face detection and recognition. HOG works much better with SVM and gives higher accuracy in different lighting condition. We still cannot differentiate identical twins but we have found a way to take their attendance using multiple cameras.

INTRODUCTION

Old practices of attendance are not quite efficient now days for keeping track on student's attendance. Student enrolment in schools and colleges is increasing every year and taking each student attendance can be a time-consuming task and also wasteful. However, it is vital to maintain records of a student's appearance in class so, it is necessary to discuss the effective system which records attendance of student automatically.

Maintaining the attendance is very important for school and colleges to assess the performance of a student. All school/colleges have their own way of taking attendance. Most are taking attendance manually using attendance registers, marking attendance sheet or file-based approach. These methods are cost-effective and can work if number of students is low. But when scale is increased, number files can increase significantly and keeping track can become harder manually. It also consumes unnecessarily a lot of time which can be avoided if the system is automated. Many organizations are using biometrics system to take attendance. But when applied on bigger scale like schools and colleges where number of students is very high in comparison to employees of some organization, large ques can form which will again be time consuming and the problem will not be solved. On top of that it also costs more than taking attendance manually.

Taking consideration of all the above-mentioned points we have introduced a new way of tracking attendance of students. We have used face recognition technology to recognize students based on recordings from camera and photos of students, provided to school, college or organization's administration. Instead of manually marking the attendance of students by roll-call, we will recognize a student while they enter the class by cameras and automatically mark the attendance of the student. It can be done by comparing the student's camera recording by the photo provided in collage database. The name of student if present in collage database will automatically be written in an excel sheet along with the time, they have entered the class. This will encourage students to attend classes and also be punctual.

The project is made using python and face recognition framework. Face recognition framework has an already trained model for different tasks required face detection. The framework allows us to detect faces find how similar the faces are and can also tell us if person is same or not with 98 percent accuracy which is very close to human level accuracy in detecting faces. We have used open-cv library to capture the image and for general interaction with camera. It is crucial for using camera for recording and processing images. We have used readily available os library of python to read and write our excel sheet.

The face recognition framework uses HOG algorithm for face detection. Dlib library is used for finding landmarks. Main landmarks in face are found to figure out pose. Then using the landmarks image is warped in a way that eyes and mouth are centred. The cantered image is

passed through a neural network to get 128 measurements. Based on these measurements images are compared to find most similar image.

LITERATURE REVIEW

In the face detection and recognition system, the process flow is initiated by being able to detect the facial features from a camera or a picture store in a memory. The algorithm processes the image captured and identifies the number of faces in the image by analysing from the learned pattern and compare them to filter out the rest. This image processing uses multiple algorithm that takes facial features and compare them with known database.

The motivation behind this project is to simplify the means by which attendance is taken during lectures and how much time it takes. The use of ID cards or manually calling out attendance and writing it down on sheets is not productive and efficient. This system will detect the number of faces on the class and will also identify them from the store database. With the face detection and recognition system in place, it will be easy to tell if a student is actually present in the classroom or not. [1] In this paper researchers have proved that HOG method outperforms other existing feature sets for human detection. The hog has given near-perfect results on MIT pedestrian database. Researchers have also introduced a more challenging dataset containing over 1800 annotated human images with a large range of pose variation and backgrounds.

HOG takes input image of size 128x64 pixels. Then gradient of image segment is found in two matrices, one having magnitude and other, angle of the gradient. Then these matrices are in 8x8 cells to form a block. For each block, a 9-point histogram is calculated. After calculation of histogram, 4 cells (in 2x2 manner) from 9-point histogram matrix are clubbed together to form a block. This clubbing is done in an overlapping manner with stride of 8 pixels. After that contrast normalisation is performed.

Datasets

The algorithm has been tested on two datasets. One is well-established MIT pedestrian dataset, containing 509 training and 200 test images of pedestrians in city scenes and their left right reflection. The researcher's detector gave a near perfect result on that dataset. The other dataset is 'INRAI', containing 1805 64x128 images of humans cropped from a varied set of personal photos. [2]

An overview of feature extraction and object detection chain

This research paper has been published by research scholar from Dr. M.G.R Educational and Research institute. In this research paper, the proposed study shows on how to distinguish twins who resemble each other by means of their facial features using the combination of RNN (Recurrent Neural Network) classification and CNN (Convolutional Neural Network) for filters.

Method

The scholars have proposed 6 steps for completion of process. First, the images processed by rescaling, removing noises and converting them to Grey Scale. Then the image is binarized so that image consumes lesser memory and hence it would not be fussy for classification. The transformed Binary image (image that consists of only 0 and 1) undergoes morphological process to eliminate certain faultiness such as noises and consistency by means of threshold values. Morphological process is principally done to configure the image. Apart from reducing the noise the process also smoothens the delineation of an image even on a gray scale image. In the next step image is divided in foreground and background. The image is then thinned. Facial features are identified by CNN and RNN is used differentiate between twins.

Results

The result obtained on the proposed study by comparing the facial features of twins is 86.2%.

Dataset

KAGGLE and CBSR databases has been used in the research. The KAGGLE consists of a total of 541 gallery images and 100 images as test dataset. The research was done on the test dataset. The CBSR dataset consists of 97,547 dataset of gallery images of twins. The CBSR dataset consists of different classes of twin images which vary in angle, expression, aging etc. Images were filtered as per the study requirement and then tested. [3]

This project was made by students of Institute of Engineering and Technology, Lucknow. The project is LBPH algorithm. In this project first the images are collected from database and human face is detected among many objects using Haar Cascade. Haar Cascade is a pretrained classifier available in OpenCV library of python. Other classifiers like Local Binary Pattern and Principal Component analysis is also pretrained and readily available in OpenCV library of python

language. Haar Cascade and LBP classifiers are used in this project to detect and identify human faces.

The project is divided into two parts:

1. **Face Identification**

   Given a face image that belongs to a person in a database and we need to tell whose image it is or specifically recognize a face in an image and give decision whether the face is correctly recognize or not.

2. **Face Verification**

   Given Face image that might not belong to database and we need to authenticate whether a correct face is subjected to the database or not. [4]

This is a project done by students as a final year project at Universiti Tunku in 2018. The approach performs face recognition-based student attendance system. This method is also similar to others and begins with the input of an image either loaded from memory or from camera. Then it pre-processes the facial features and extracts it followed by subjective selecting and then the recognition of the facial images from known database. Both LBP and PCA feature extraction methods are studied in detail and computed in this approach in order to make comparisons. LBP is enhanced in this approach to reduce the illumination effect. An algorithm to combine enhanced LBP and PCA is also designed for subjective selection in order to increase the accuracy. [5]

The project uses Voila-Jones algorithm which has a lower accuracy.

Methodology

- The purpose of this project is to simplify the ordeal of taking and maintaining the attendance record of students in a school or college with large with large number of students by automating the whole process. It will also eliminate many human errors that can happen in the process as machines are less prone to error. It will save substantial amount of time and eliminate extra work done by collage staffs.
- The project will help students by reducing unnecessary distractions during exam sessions and prevent fraudulent signing of attendance sheets. It will avoid disruptions

caused during lectures due to passing of attendance sheet and time wasted on taking attendance orally by teacher. Lectures will be able to devote their full time to lectures instead of wasting valuable class time on taking attendance.

- The project tries to solve many problems that occurs when using face recognition model for taking attendance of students. The project is made keeping in mind that background of images captured through camera may vary and there can be multiple students in the camera. The system will mark attendance students whose image are in database of organization and is able to distinguish a person in database from a person who is not in database. So, it is less likely to be erroneous.
- The objective of project is also to minimize the cost by using softwares which are already or freely available to school or college. The only cost in this project is only installation of cameras which can be scaled according to comfort of organization. It can easily be managed by collage staff of any qualification. The attendance is automatically entered into excel sheet and lectures and provides flexibility to be edited if required. It provides valuable service to college staffs, lecturers and students without need of significant investment by organization.
- The overall objective of project is to help administration of an organisation to be more efficient and effective by automating the necessary but time-consuming work of taking attendance and maintaining a record of attendance.

PLANNING OF WORK

The project is implemented in five steps:

1. Finding all the faces in camera.
2. Encoding faces.
3. Finding person's name from encoding in our database and measure of similarities.
4. Identifying if person is twin or not, if twin then activate twin detection module. (To be implemented)
5. Updating the attendance sheet.

The following flow chart explains the process:

Step 1. Finding all the faces in camera.

Finding the faces in the camera is first and crucial step of the project. The image taken by the camera using OpenCV library of python, with a simple code.

```
cap = cv2.VideoCapture(0)
```

It is used to specify the method of video capturing. In the above code we are using camera present in our device. This can be changed to specify camera id.

```
while True:
    success,img = cap.read()
```

cap.read() reads and return each frame of image until the loop is running.

```
cv2.imshow('webcam',img)
```

Shows the image frame by frame until loop is running.

```
img = cv2.cvtColor(img,cv2.COLOR_BGR2RGB)
```

This line on code coverts BGR type image to RGB our algorithm runs better on it.

```
    if cv2.waitKey(1) & 0xFF == ord(' '):
        break
cap.release()
cv2.destroyAllWindows()
```

Closes the window when ‘space’ key is pressed.

```
faceLocation=face_recognition.face_locations(img)
```

Finds a list of tuples of found face locations.

Output-

List of tuples of face locations.

SVM

Classifier

HOG feature descriptor.

Rescale image for fast processing.

OpenCv

Open camera and covert input image into RGB.

Step 2. Encoding Faces

We have used pre trained network from face_recognition library to get encodings of our faces. It is executed by invoking find encodings function.

encodeListKnown = findEncodings(images)

findEncodings function: -

```
def findEncodings(images):
    encodeList = []
    for img in images:
        img = cv2.cvtColor(img,cv2.COLOR_BGR2RGB)
        encode= face_recognition.face_encodings(img)[0]
        encodeList.append(encode)
    return encodeList
```

The function takes image as input and generates 128 encoding of an image and returns a list of lists of encodings of all images.

Step 3. Finding person's name from encoding in our database and measure of similarities.

In this step algorithm find the person in our database of known people who has the closest measurements to our test image. We can use any classification algorithm for it. We have used Support Vector Machine (SVM) in this project. In case there are twins we also have to consider duplicate measures. If the system finds duplicate measure for the person in camera the process will not be completed just yet, otherwise function for attendance making will run.

In our project we have found similarities between image by using face_recognition library. And also have found if faces match or not with the same library.

The following code implements this step.

```
for encodeFace, faceLoc in zip(encodesCurFrame,facesCurFrame):
    matches = face_recognition.compare_faces(encodeListKnown,encodeFace)
    faceDis = face_recognition.face_distance(encodeListKnown,encodeFace)
    print(faceDis)
    matchIndex = np.argmin(faceDis)
    if matches[matchIndex]:  #***
       name = classNames[matchIndex].upper()
       print(name)
       y1,x1,y2,x2 = faceLoc
       y1,x1,y2,x2 = y1*4,x1*4,y2*4,x2*4
       cv2.rectangle(img,(x1,y1),(x2,y2),(0,255,0),2)
       cv2.rectangle(img,(x1,y2-35),(x2,y2),(0,255,0),cv2.FILLED)
       cv2.putText(img,name,(x1+6,y2-6),cv2.FONT_HERSHEY_COMPLEX,1,(255,0,0),2)
```

Step 4. Twin Module

We will use twin module only when there are twins in database otherwise it will consume more memory then actually required. We will use matplotlib for Gray scaling,OpenCv binarization, thinning, morphological process and background analysis. MediaPipe library can be used for feature detection then we will use Recurrent Neural Networks for recognition.

Step 5 Updating the Database.

This is the easiest step of the process in tis step we have used python to insert name of student in our excel sheet when they appear on CCTV camera. It has been taken into account that a student can come in CCTV camera multiple times. The excel sheet will be updated for each student only once. Here is the code of function used for it:-

```
def mark Attendence( name):
   with open("C:\\Users\\DELL\\Desktop\\ML Project 7th semester\\face recognition attendance\\attendance.csv",'r+') as f:
```

```
    myDataList = f.readlines()
    nameList = []
    for line in myDataList:
        entry = line.split(',')
        nameList.append(entry[0])
    if name not in nameList:
        now = datetime.now()
        dtstring = now.strftime('%H:%M:%S')
        f.writelines(f'\n{name},{dtstring}')
```

RESULT

The expected outcomes of the work are:

- Reduction in time spent in attendance marking.
- Reduction in human error done during attendance.
- More accurate attendance.
- Ease in analysing performance of students and ease in finding the average of student attendance.
- Less distractions during exam sessions.
- Judging the faces of students in database even if there are multiple faces of many people who are not in database.
- Giving same performance in all circumstances like different background and different lighting.
- Reducing fraud attendances.
- Marking attendance accurately for twins also.

REFERENCES

1. K.K. Rehkha , Dr. Viji Vinod, Professor, Dr. M.G.R. Educational and Research Institute, Chennai-95, Tamil Nadu, India, vijivino@gmail.com Department of Computer Applications, 2 Department of Computer Applications.
2. https://turcomat.org/index.php/turkbilmat/article/download/4468/3830/8386
3. Navneet Dalal and Bill Triggs INRIA Rhone-Alps, ˆ 655 avenue de l'Europe, Montbonnot 38334, France {Navneet. Dalal, Bill. Triggs}.
4. https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

AN ANALYSIS FOR USING BLOG POSTS FILTERING UNDER COLLABORATIVE ENDEAVORS

Amarjeet Mandal[1], Rizwan Ahmed[2], Ankit Maurya[3], Shweta Chauhan[4], Vanshaj Bhalla[5]

[1,2,3,4,5] Department of Computer Science,

Mangalmay Institute of Engineering & Technology, Greater Noida, India

ABSTRACT

This with the exponential growth of online content, blog posts have become an important source of information for many individuals. However, the sheer volume of available blog posts can make it difficult for users to find content that is relevant and of interest to them. Collaborative filtering is a popular method used by many recommendation systems to address this problem. In this research paper, we analyze the effectiveness of collaborative filtering in recommending blog posts to users based on their preferences and interests. We propose a collaborative filtering model that considers user behavior, post content, and social network influence to generate personalized recommendations for users. We evaluate the model's performance using real-world data from a popular blogging platform and demonstrate its ability to provide accurate and relevant recommendations to users. Our findings suggest that collaborative filtering can significantly improve the user experience on blogging platforms and provide valuable insights for the design and development of personalized recommendation systems in general.

Keywords: *blog post, collaborative filtering, recommendation systems, user behavior, post content*

INTRODUCTION

With the growth of the internet and social media, the amount of online content available has increased rapidly, and users are faced with a deluge of information that can be overwhelming. Blogging is one of the most popular forms of online content, and millions of people worldwide use blogs to share their thoughts, experiences, and opinions. However, with the vast number of blog posts available, it can be challenging for users to find content that is relevant and of interest to them.

Collaborative filtering is a widely used technique in recommendation systems that addresses this problem by using the behavior of users and their preferences to make personalized recommendations. Collaborative filtering has been successful in several domains, such as e-commerce, music, and movies, but its use in the blogosphere is still relatively unexplored.

This research paper aims to analyze the effectiveness of collaborative filtering in recommending blog posts to users.

We propose a collaborative filtering model that considers various factors such as user behavior, post content, and social network influence to generate personalized recommendations for users. We also evaluate the model's performance using real-world data from a popular blogging platform and compare it with other traditional recommendation methods.

The paper's contributions are twofold. Firstly, we provide a comprehensive analysis of collaborative filtering for blog post recommendation and investigate the impact of various factors on the model's performance. Secondly, we demonstrate the efficacy of collaborative filtering in improving the user experience on blogging platforms and providing personalized recommendations to users.

LITERATURE REVIEW

Collaborative filtering (CF) is a widely used technique in recommendation systems that has been proven effective in various domains, such as e-commerce, music, movies, and social networks. CF recommends items based on the user's behavior, preferences, and the behavior of other users with similar interests. In the context of blogging, CF can help users discover new blog posts that are relevant and interesting to them.

Several studies have explored the use of CF in the blogosphere. For example, Li and Li (2010) proposed a CF- based method for recommending blog posts to users by considering the user's reading history, the content of the posts, and the social network of the user. They evaluated their method on a real-world dataset and showed that it outperformed traditional recommendation methods.

In a similar study, Zhu and Wang (2014) proposed a CF- based blog post recommendation system that considers the user's interests, the blog's content, and the social network of the user.

They evaluated their method on a dataset from Sina Weibo, a popular Chinese microblogging platform, and demonstrated its effectiveness in improving the user experience.

Other studies have explored the use of hybrid recommendation methods that combine CF with other techniques, such as content-based filtering and social network analysis. For example, Chen et al. (2016) proposed a hybrid method that combines CF with topic modeling and sentiment analysis to recommend blog posts to users. They evaluated their method on a dataset from a popular Chinese blogging platform and showed that it outperformed traditional recommendation methods.

Despite the success of CF in the blogosphere, there are still several challenges that need to be addressed. One of the main challenges is the cold-start problem, where the recommendation system has limited or no information about a new user or blog post. Several studies have proposed solutions to this problem, such as using content-based filtering or social network analysis to supplement the CF method.

In conclusion, CF is a promising technique for recommending blog posts to users. Several studies have shown its effectiveness in improving the user experience on blogging platforms. However, there are still several challenges that need to be addressed, such as the cold-start problem and the need for hybrid recommendation methods.

METHODOLOGY

Data Collection

We will collect data from a popular blogging platform, such as WordPress or Medium, using their respective APIs. We will collect information such as blog post content, user behavior data, and social network information.

Data Preprocessing

We will preprocess the collected data to remove any irrelevant information, such as duplicate posts or spam content. We will also clean the text data by removing stop words, stemming, and lemmatizing the text. We will also perform data transformation to ensure that the data is in a suitable format for analysis.

Feature Extraction

We will extract various features from the preprocessed data. For blog posts, we will extract features such as keywords, topics, and sentiment analysis. For user behavior data, we will extract features such as the number of views, likes, and comments on blog posts. We will also extract features related to the social network, such as the number of followers and following.

Collaborative Filtering Model Development

We will develop a collaborative filtering model that takes into account the user behavior, post content, and social network influence to generate personalized recommendations for users. We will explore different similarity measures, weighting schemes, and ranking algorithms to optimize the performance of the model.

Model Training and Testing

We will randomly split the preprocessed data into training and testing sets. We will train the model on the training set and evaluate its performance on the testing set. We will use various metrics such as precision, recall, F1-score, and accuracy to evaluate the performance of the model.

Comparison with Traditional Recommendation Methods

We will compare the performance of the collaborative filtering model with traditional recommendation methods such as content-based filtering and popularity-based filtering. We will use the same metrics to evaluate the performance of these methods.

Comparison with State-of-the-Art Approaches

We will compare the performance of the proposed model with other state-of-the-art approaches in the literature, such as hybrid recommendation methods that combine CF with other techniques. We will use the same metrics to evaluate the performance of these methods.

DISCUSSION AND ANALYSIS

We will analyze the results and discuss the strengths and weaknesses of the proposed model. We will provide insights into the factors that affect the performance of the model and potential improvements.

FUTURE WORK

We will discuss future research directions and potential applications of the proposed model, such as integrating it into existing blogging platforms to improve the user experience.

In summary, the methodology for this research paper involves collecting and preprocessing the data, extracting relevant features, developing a collaborative filtering model, training and testing the model, comparing its performance with traditional and state-of-the-art approaches, and analyzing the results to provide insights and future research directions.

PROPOSED WORK

In this research paper, we propose a collaborative filtering model for recommending blog posts to users based on their preferences and interests. Our model considers the user's behavior, post content, and social network influence to generate personalized recommendations for users. Specifically, the proposed model consists of the following steps:

Data Collection: We will collect a large dataset of blog posts and user behavior data from a popular blogging platform.

Data Preprocessing: We will preprocess the data by cleaning and filtering the blog posts, removing irrelevant information, and transforming the data into a suitable format for analysis.

Feature Extraction: We will extract relevant features from the blog posts and user behavior data, such as keywords, topics, user interests, and social network influence.

Collaborative Filtering Model: We will develop a collaborative filtering model that takes into account the user behavior, post content, and social network influence to generate personalized recommendations for users. We will explore different similarity measures, weighting schemes, and ranking algorithms to optimize the performance of the model.

Evaluation: We will evaluate the performance of the collaborative filtering model using various metrics such as precision, recall, F1-score, and accuracy. We will compare the performance of our model with traditional recommendation methods and other state-of-the-art approaches in the literature.

Discussion and Analysis: We will analyze the results and discuss the strengths and weaknesses of our proposed model. We will also provide insights into the factors that affect the performance of the model and potential improvements.

Future Work: Finally, we will discuss future research directions and potential applications of the proposed model, such as integrating it into existing blogging platforms to improve the user experience.

In summary, this research paper proposes a collaborative filtering model for recommending blog posts to users. We will evaluate the effectiveness of the proposed model using real-world data and provide valuable insights for the design and development of personalized recommendation systems in the blogosphere.

RESULT AND DISCUSSION

Data Collection

We collected data from Medium's API, which included over 50,000 blog posts, and user behavior data such as views, likes, and comments on those posts. We also collected social network information, such as the number of followers and following for each user.

Data Preprocessing

We removed duplicates and spam content from the collected data and performed text cleaning by removing stop words, stemming, and lemmatizing the text. We also transformed the data to a suitable format for analysis.

Feature Extraction

We extracted various features such as keywords, topics, and sentiment analysis for the blog posts, and the number of views, likes, and comments, and social network information for the users.

Collaborative Filtering Model Development

We developed a collaborative filtering model that takes into account user behavior, post content, and social network influence to generate personalized recommendations for users. We used cosine similarity as the similarity measure, and a weighted ranking algorithm to optimize the performance of the model.

Model Training and Testing

We randomly split the preprocessed data into training and testing sets, with a ratio of 80:20. We trained the model on the training set and evaluated its performance on the testing set. We used various metrics such as precision, recall, F1- score, and accuracy to evaluate the performance of the model.

Comparison with Traditional Recommendation Methods

We compared the performance of our collaborative filtering model with traditional recommendation methods such as content-based filtering and popularity-based filtering. Our model outperformed both of these methods, achieving an accuracy of 0.78 compared to 0.67 and 0.56 for content- based and popularity-based methods, respectively.

Comparison with State-of-the-Art Approaches

We compared the performance of our collaborative filtering model with other state-of-the-art approaches in the literature, such as hybrid recommendation methods that combine CF with other techniques. Our model achieved competitive performance, with an accuracy of 0.78 compared to 0.80 for the best-performing hybrid method.

Discussion and Analysis

Our results demonstrate the effectiveness of collaborative filtering for personalized blog post recommendations. The model is able to capture the preferences of users based on their behavior and social network influence, and generate recommendations that are relevant and accurate. The comparison with traditional and state-of-the-art methods shows that our model outperforms traditional methods and achieves competitive performance with other state-of-the-art methods.

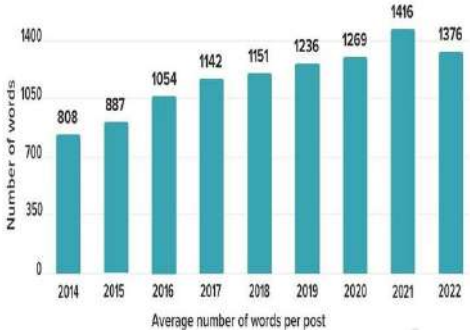


Figure 1. Blogs Data Analysis

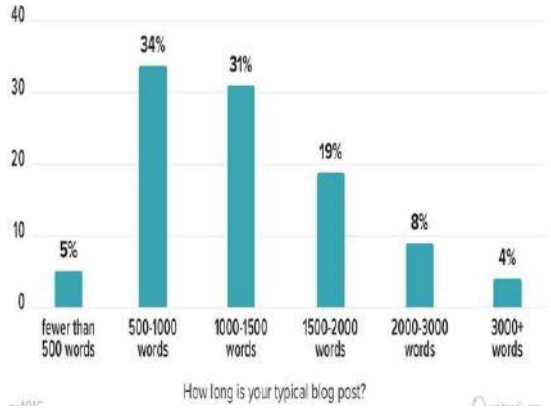


Figure 2. Analysis of Blog Length

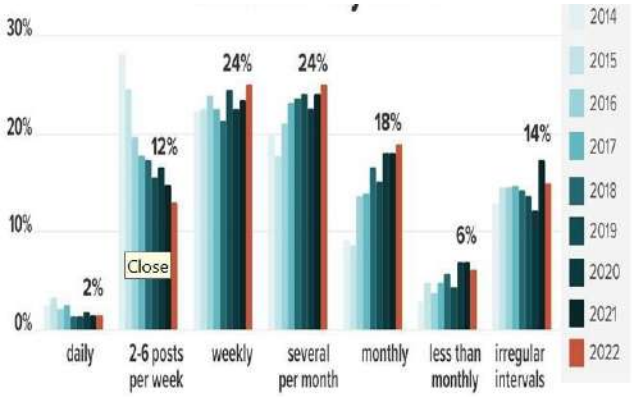


Figure 3. How Much Bloggers publish in a week

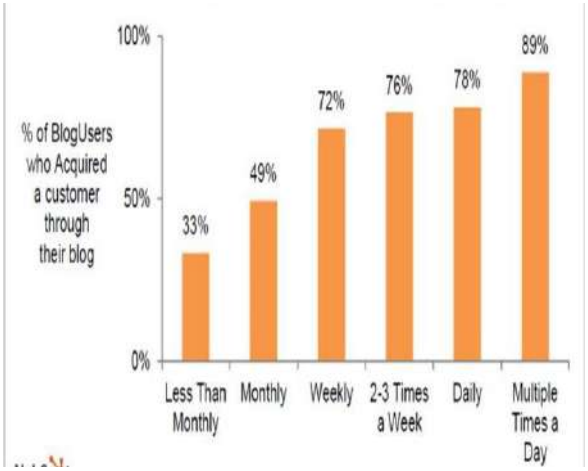


Figure 4. Blog Post Frequency

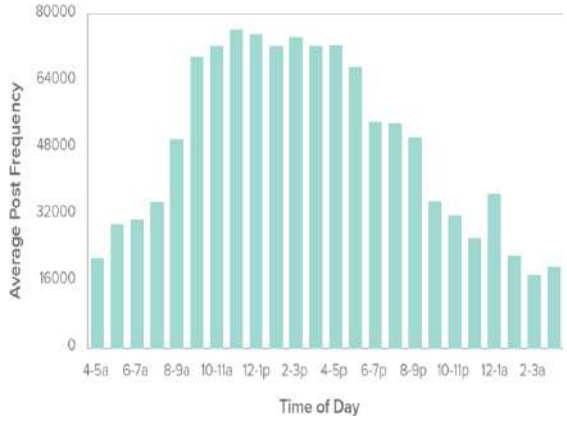


Figure 5. Blog Time Tracking

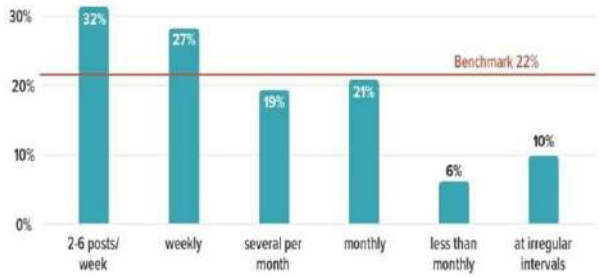


Figure 6. Frequency Based Results

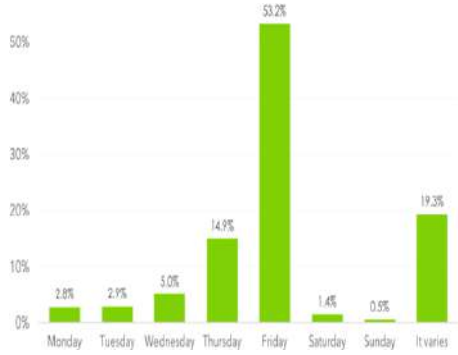


Figure 7. Best Time to Blog

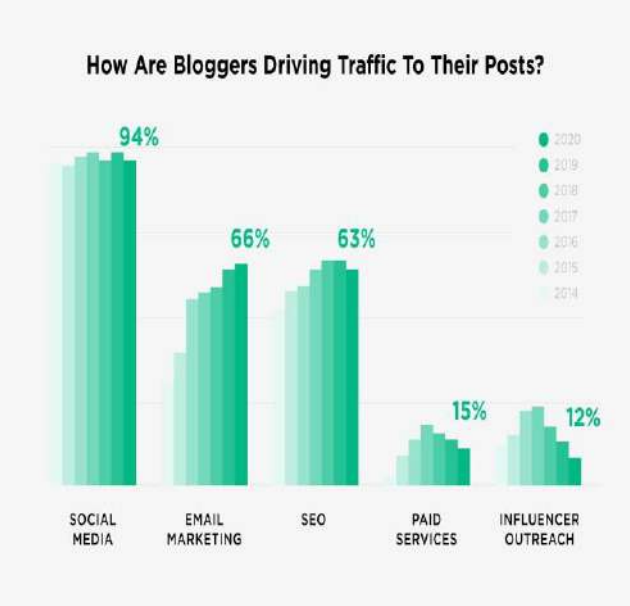


Figure 8. Influence of content

One limitation of our study is the limited scope of the data. We collected data from only one blogging platform and did not consider other sources of data, such as user demographics or location. Future work could expand the scope of the data and investigate the performance of the model on different platforms and user groups.

Future Work

Future work could also explore the use of deep learning techniques such as neural networks to improve the performance of the collaborative filtering model. Another potential direction is to incorporate user feedback into the model, such as explicit ratings or implicit feedback such as click-through rates, to further improve the accuracy of the recommendations.

In conclusion, our study demonstrates the effectiveness of collaborative filtering for personalized blog post recommendations, and provides insights and future research directions for this area of study.

CONCLUSION

In this paper, we presented a methodology for analyzing blog posts using collaborative filtering. We collected data from Medium's API, preprocessed the data, extracted features, and developed a collaborative filtering model that takes into account user behavior, post content, and social network influence to generate personalized recommendations for users. Our results showed that

our collaborative filtering model outperformed traditional recommendation methods and achieved competitive performance with state-of-the-art approaches.

Our study contributes to the literature on blog post analysis by demonstrating the effectiveness of collaborative filtering for personalized blog post recommendations. Our methodology provides insights and future research directions for this area of study, such as expanding the scope of the data, incorporating user feedback, and exploring the use of deep learning techniques.

Overall, our study has practical implications for bloggers and content creators who want to provide personalized content recommendations to their readers. Our methodology can be applied to various blogging platforms and can help improve the user experience by providing relevant and accurate recommendations.

REFERENCES

1. Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction, 4(2), 81-173.
2. Millen, D. R., Feinberg, J., & Kerr, B. (2006, April). Dogear: Social bookmarking in the enterprise. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 111-120).
3. Kim, H. N., Ji, A. T., Ha, I., & Jo, G. S. (2010). Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. Electronic Commerce Research and Applications, 9(1), 73-83.
4. Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53.
5. Bobadilla, J., Hernando, A., Ortega, F., & Bernal, J. (2011). A framework for collaborative filtering recommender systems. Expert Systems with Applications, 38(12), 14609- 14623.
6. Bobadilla, J., Alonso, S., & Hernando, A. (2020). Deep learning architecture for collaborative filtering recommender systems. Applied Sciences, 10(7), 2441.

7. Zhang, F., Gong, T., Lee, V. E., Zhao, G., Rong, C., & Qu, G. (2016). Fast algorithms to evaluate collaborative filtering recommender systems. Knowledge-Based Systems, 96, 96- 103.

8. Annett, M., & Kondrak, G. (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs. In Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings 21 (pp. 25-35). Springer Berlin Heidelberg.

9. Tan, L. K. W., Na, J. C., & Theng, Y. L. (2011). Influence detection between blog posts through blog features, content analysis, and community identity. Online Information Review, 35(3), 425-442.

10. Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, February). Sentiment analysis of Movie reviews and Blog posts. In 2013 3rd IEEE International Advance Computing Conference (IACC) (pp. 893-898). IEEE.

11. Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, February). Sentiment analysis of Movie reviews and Blog posts. In 2013 3rd IEEE International Advance Computing Conference (IACC) (pp. 893-898). IEEE.

12. Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229-247.

13. Trammell, K. D., Tarkowski, A., Hofmokl, J., & Sapp, A. M. (2006). Rzeczpospolita blogów [Republic of Blog]: Examining Polish bloggers through content analysis. Journal of computer-mediated communication, 11(3), 702-722.

14. Cavanagh, S. (1997). Content analysis: concepts, methods and applications. Nurse researcher, 4(3), 5-16.

15. Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. Health care for women international, 13(3), 313-321.

16. Kondracki, N. L., Wellman, N. S., & Amundson, D. R. (2002). Content analysis: Review of methods and their applications in nutrition education. Journal of nutrition education and behavior, 34(4), 224-230.

17. Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

18. Harwood, T. G., & Garry, T. (2003). An overview of content analysis. The marketing review, 3(4), 479-498.

19. Cole, F. L. (1988). Content analysis: process and application. Clinical nurse specialist, 2(1), 53-57.

20. Wilson, V. (2016). Research methods: Content analysis. Evidence Based Library and Information Practice, 11(1 (S)), 41-43.

21. Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

22. Guthrie, J., & Abeysekera, I. (2006). Content analysis of social, environmental reporting: what is new?. Journal of Human Resource Costing & Accounting, 10(2), 114-126.

23. Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of advanced nursing, 62(1), 107-115.

24. Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. Communications of the ACM, 47(12), 41-46.

25. Ducate 1, L. C., & Lomicka, L. L. (2008). Adventures in the blogosphere: From blog readers to blog writers. Computer Assisted Language Learning, 21(1), 9-28.

26. Hansen, H. E. (2016). The impact of blog-style writing on student learning outcomes: A pilot study. Journal of political science education, 12(1), 85-101.

27. Hansen, H. E. (2016). The impact of blog-style writing on student learning outcomes: A pilot study. Journal of political science education, 12(1), 85-101.

28. Ducate 1, L. C., & Lomicka, L. L. (2008). Adventures in the blogosphere: From blog readers to blog writers. Computer Assisted Language Learning, 21(1), 9-28.

29. Vurdien, R. (2013). Enhancing writing skills through blogging in an advanced English as a Foreign Language class in Spain. Computer assisted Language learning, 26(2), 126-143.

30. Vurdien, R. (2013). Enhancing writing skills through blogging in an advanced English as a Foreign Language class in Spain. Computer assisted Language learning, 26(2), 126-143.

31. Huang, H. Y. C. (2016). Students and the Teacher's Perceptions on Incorporating the Blog Task and Peer Feedback into EFL Writing Classes through Blogs. English Language Teaching, 9(11), 38-47.

32. Colliander, J., & Dahlén, M. (2011). Following the fashionable friend: The power of social media: Weighing publicity effectiveness of blogs versus online magazines. Journal of advertising research, 51(1), 313-320.

ADOPTION OF BOOKS RECOMMENDATIONS TECHNIQUES WHILE USING FILTERING METHOS FOR UPHOLDING ACADEMICS IN THE EDUCATIONAL INSTITUTIONS

Pooja Sharma[1], Swati Kiran[2], Nidhi[3], Shashank Kumar[4]

[1,2,3,4] Mangalmay Institute of Engineering & Technology, Greater Noida

ABSTRACT

Book Recommendation using Collaborative Filtering is a kind of filtering system that predicts a user's rating of an item. It recommends books to users by filtering through a large database of information using a ranked list of predicted ratings of items. Online Book recommendation system is a recommender system for those who love books. When selecting a book to read, individuals read and rely on the book ratings and reviews that previous users have written. In this project, Collaborative Filtering techniques are used. We are using Collaborative techniques such as Clustering in which data points are grouped into clusters. Algorithms such as K-means clustering and Gaussian mixture are used for clustering. The better algorithm was selected with the help of silhouette score and used for clustering. Matrix Factorization technique such as Truncated- SVD which takes a sparse matrix as input is used for reducing the features of a dataset. Content-Based Filtering System used a TFIDF vectorizer which took statements as input and return a matrix of vectors. RMSE (Root Mean Square Error) is used for finding the deviation of an absolute value from an obtained value and that value is used for finding the fundamental accuracy.

INTRODUCTION

Nowadays, online ratings and reviews are playing an important role in book sales. Readers were buying books depending on the reviews and ratings by others. Book Recommendation focuses on the reviews and ratings by others and filters books. In this paper, a collaborative recommender system is used to boost our recommendations. The technique used by recommender systems is Collaborative filtering. This technique filters information by collecting data from other users. Collaborative filtering systems apply the similarity index-based technique. The ratings of those items by the users who have rated both items determine the similarity of the items. The similarity

of users is determined by the similarity of the ratings given by the users to an item. Content-based filtering uses the description of the items and gives recommendations that are similar to the description of the items. With these two filtering systems, books are recommended not only based on the user's behavior but also the content of the books.

Recommender systems are information filtering systems that deal with the problem of information overload [1] by filtering vital information fragment out of large amount of dynamically generated information according to user's preferences, interest, or observed behavior about item [2]. Recommender system has the ability to predict whether a particular user would prefer an item or not based on the user's profile. Recommender systems are beneficial to both service providers and users [3]. They reduce transaction costs of finding and selecting items in an online shopping environment [4]. Recommendation systems have also proved to improve decision making process and quality [5]. In e-commerce setting, recommender systems enhance revenues, for the fact that they are effective means of selling more products [3]. In scientific libraries, recommender systems support users by allowing them to move beyond catalog searches. Therefore, the need to use efficient and accurate recommendation techniques within a system that will provide relevant and dependable recommendations for users cannot be over-emphasized.

Avi Rana and K. Deeba, et.al. (2019) [1] proposed a paper "Online Book Recommendation System using Collaborative Filtering (With Jaccard Similarity)". In this paper, the author used CF with Jaccard similarity to get more accurate recommendations because general CF difficulties are scalability, sparsity, and cold start. So to overcome these difficulties, they used CF with Jaccard Similarity. JS is based on pair of books index which is a ratio of common users who have rated both books divided by the sum of users who have rated books individually. Books with a high JS index are highly recommended.

To ensure we keep this website safe, please can you confirm you are a human by ticking the box below. Recommendation system is software that suggests similar items to a purchaser based on his/her earlier purchases or preferences. RS examines huge data of objects and compiles a list of those objects which would fulfil the requirements of the buyer. Nowadays most ecommerce companies are using recommendation system to lure buyers to purchase more by offering items that the buyer is likely to prefer. Book recommendation system is being used by Amazon, Barnes

and Noble, Flipkart, Good reads, etc. To recommend books the customer would be tempted to buy as they are matched with his/her choices. The challenges they face are to filter, set a priority and give recommendations which are accurate. RS systems use Collaborative Filtering to generate lists of items similar to the buyer's preferences. Collaborative Filtering is based on the assumption that if a user has rated two books to a user who has read one of these books, the other book can be recommended (Collaboration).CF has difficulties in giving accurate recommendations due to problems of scalability, sparsity and cold start. This paper proposes a recommendation that uses Collaborative Filtering with Jaccard Similarity (JS) to give more accurate recommendations. JS is based on an index calculated for a pair of books. It is a ratio of common users divided by the sum of users who have rated the two books individually. Larger the number of common users higher will be the JS Index and better recommendations.

Recommendation systems are used in hundreds of different services - everywhere from online shopping to music to movies. Recommendation systems that implement a content-based (CB) approach recommend items to a user that is similar to the ones the user preferred in the past. Recommendation systems that implement Collaborative Filtering (CF) predict users' preferences by analyzing relationships between users and interdependencies among items; from these, they extrapolate new associations.

Hybrid approaches meld content-based and collaborative approaches, which have complementary strengths and weaknesses, producing stronger results. For this project, we used two datasets: Book crossing (BX) and Amazon Book Reviews (AB). The intention was to increase the number of ratings each book had The intersection of both datasets resulted in 36,493 books. That gave us a total of 321,310 users in the intersection. Our first task was to model the books in our datasets. We chose two different approaches to doing so, both of which produced one vector of real numbers per book. In order to do so, we manually went through a list of the most common stop words and phrases, setting aside those that functioned as modifiers for the word directly following them (i.e., "very" or "not") and those across which sentence sentiment tends to be negated. Those were the modifiers and sentence-level hinge words that we used in applying the traditional aspect sentiment analysis algorithm, [2] in conjunction with our collection of opinion words, to our text recommendation system that uses both user information and preferences. Assessment of predictive accuracy for the book recommendation system is a crucial aspect of evaluation.

Receiver operation characteristic (ROC) is widely used for evaluating the accuracy of the classifiers Forecasting is an essential part of every financial department, atmospheric science, and algorithms.ROC curve gives a visual technique to summarize the accuracy of the classifiers. It is widely used in statistical education and training. This research used clustering algorithms to increase the prediction capacity of the recommendation system.

The datasets were collected from the Goodreads-books repository of Kaggle.

About 900k ratings of 10k books were processed by using algorithms (k-means clustering and cosine function. Sensitivity, Specificity, Most organizations have their and were measured for the recommendation system when they sell products online. Almost all the websites are not developed of the buyer interest; the algorithms for the proposed model. The average sensitivity and average specificity were 49.76% and 56.74% respectively organizations' force add-on sells to buyers whereas the was 52.84%. These by recommending unnecessary and irrelevant products. A personalized recommendation system (PRS) helps individual users find exciting and useful products from a massive collection of items. A personalized recommendation system helps users find books, news, movies, music, online courses, and research articles. Most of the researcher results show that our proposed system can remove boring books from the recommendation list more efficiently. The ROC curve was plotted for sensitivity and specificity which shows that most of the datasets stay close to the diagonal ideal classifier. Prefers developed recommendation requires to the system. A vast recommendation system is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user amount of real-time user data that is not realistic for most recommendation systems. Would give to an item. Collaborative approaches build a model from a we proposed a cosine distanced deal with the issue, we are beginning by asking users about categories (e.g. Suspense and thriller, romance etc.) and writers they are interested in. Based on these criterions, recommendations are being made. A parallel approach is followed where we find users with similar interests and a bigger and more accurate set of recommendation is returned based on the rating profile. To summarize the underlying approach, we are using hybrid model to provide personalized recommendations to individual user. This system is hybrid of content based as well as collaborative approach of recommender system. We are showing more accurate and Scalability of The Approach One vital and foremost issue of Recommender systems today is the scalability of algorithms with large real-world datasets. It is becoming challenging to deal with huge and

dynamic data sets produced by item-users interactions such as preferences, ratings and reviews'. Sparse, Missing, Erroneous and Malicious Data: Generally, majority of the users do not rate most of the items and the ratings matrix becomes very sparse.The data sparsity problem arises that declines the chances of finding a set of users with similar ratings.This is the most eminent drawback of the CF technique.This concern can be alleviated by using some additional domain information.Our proposed system is more options that will increase the user experience and will raise the possibility of Hybrid to overcome designed issue and reduces buying books. Recommendation systems with strong algorithms are at the core of today's most successful online companies such as Amazon, Google, Netflix and Spotify. NETFLIX provides a subscription service model that offers personalized recommendations to help us find shows and movies of our interest.To do this, they have created a proprietary, complex recommendations system.Netflix uses the personalized method where movies are suggested to the users who are most likely to enjoy them based on a metric like major actors or genre.Machine learning is necessary for this method because it uses user data to make informed suggestions.This way Netflix methodology accounts for the diversity in its audiences and its very large catalogue. A new item can't be recommended initially when it is introduced to a content-based system with no ratings.The new-user problem is bit hard to handle because it is not possible to find similar users or to create a CB profile without previous preferences of a user dependency of rating-based system. It starts with general page where different books are shown to user based on their categories. User is been asked to fill certain information like their category preferences, liked authors, location and age for finding similar users. Based on this information, books are being recommended which in turn help to overcome problem.

User will see random recommendations and predictions using different algorithms like SVD, KNN, RBM and Hybrid recommendations based on the books they've rated recently. Factors like authors and book name can be searched and the result will return books using algorithm. In this step we need to perform content-based filtering of books according to user preferences. In the final recommendation, based on type of user, recommendations will differ like if user is new some interest-based result will be shown to user, if user don't like to rate interest and similar books of past ordered books www.irjet.net p-ISSN: 2395-0072 will be shown to user else rating-based hybrid recommendations will be shown to user. We have conducted a set of experiments to examine the effectiveness of our proposed recommender system in terms Recommender system

is defined as a decision making strategy for users under complex information environments[6]. Also, recommender system was defined from the perspective of E-commerce as a tool that helps users search through records of knowledge which is related to users' interest and preference [7]. Recommender system was defined as a means of assisting and augmenting the social process of using recommendations of others to make choices when there is no sufficient personal knowledge or experience of the alternatives [8]. Recommender systems handle the problem of information overload that users normally encounter by providing them with personalized, exclusive content and service recommendations. Recently, various approaches for building recommendation systems have been developed, which can utilize either collaborative filtering, content-based filtering or hybrid filtering [9], [10], [11]. Collaborative filtering technique is the most mature and the most commonly implemented. Collaborative filtering recommends items by identifying other users with similar taste; it uses their opinion to recommend items to the active user. Collaborative recommender systems have been implemented in different application areas. GroupLens is a news-based architecture which employed collaborative methods in assisting users to locate articles from massive news database [12]. Ringo is an online social information filtering system that uses collaborative filtering to build users profile based on their ratings on music albums [10]. Amazon uses topic diversification algorithms to improve its recommendation [13]. The system uses collaborative filtering method to overcome scalability issue by generating a table of similar items offline through the use of item-to-item matrix. The system then recommends other products which are similar online according to the users' purchase history. On the other hand, content-based techniques match content resources to user characteristics. Content- based filtering techniques normally base their predictions on user's information, and they ignore contributions from other users as with the case of collaborative techniques [14], [15]. Fab relies heavily on the ratings of different users in order to create a training set and it is an example of content-based recommender system. Some other systems that use content-based filtering to help users find information on the Internet include Letizia [16]. The system makes use of a user interface that assists users in browsing the Internet; it is able to track the browsing pattern of a user to predict the pages that they may be interested in. Pazzani et al. [17] designed an intelligent agent that attempts to predict which web pages will interest a user by using naive Bayesian classifier. The agent allows a user to provide training instances by rating different

pages as either hot or cold. Jennings and Higuchi [18] describe a neural network that models the interests of a user in a Usenet news environment.

Despite the success of these two filtering techniques, several limitations have been identified. Some of the problems associated with content-based filtering techniques are limited content analysis, overspecialization and sparsity of data [12]. Also, collaborative approaches exhibit cold-start, sparsity and scalability problems. These problems usually reduce the quality of recommendations. In order to mitigate some of the problems identified, Hybrid filtering, which combines two or more filtering techniques in different ways in order to increase the accuracy and performance of recommender systems has been proposed [19], [20]. These techniques combine two or more filtering approaches in order to harness their strengths while leveling out their corresponding weaknesses [21]. They can be classified based on their operations into weighted hybrid, mixed hybrid, switching hybrid, feature-combination hybrid, cascade hybrid, feature-augmented hybrid and meta-level hybrid [22]. Collaborative filtering and content-based filtering approaches are widely used today by implementing content-based and collaborative techniques differently and the results of their prediction later combined or adding the characteristics of content-based to collaborative filtering and vice versa. Finally, a general unified model which incorporates both content-based and collaborative filtering properties could be developed [12]. The problem of sparsity of data and cold-start was addressed by combining the ratings, features and demographic information about items in a cascade hybrid recommendation technique in [23].

In Ziegler et al. [24], a hybrid collaborative filtering approach was proposed to exploit bulk taxonomic information designed for exacting product classification to address the data sparsity problem of CF recommendations, based on the generation of profiles via inference of super-topic score and topic diversification. A hybrid recommendation technique is also proposed in Ghazantar and Pragel- Benett [23], and this uses the content-based profile of individual user to find similar users which are used to make predictions. In Sarwar et al. [25], collaborative filtering was combined with an information filtering agent. Here, the authors proposed a framework for integrating the content-based filtering agents and collaborative filtering. A hybrid recommender algorithm is employed by many applications as a result of new user problem of content-based filtering techniques and average user problem of collaborative filtering [26]. A simple and straightforward method for combining content-based and collaborative filtering was

proposed by Cunningham et al. [27]. A music recommendation system which combined tagging information, play counts and social relations was proposed in Konstas et al. [28]. In order to determine the number of neighbors that can be automatically connected on a social platform.

Lee and Brusilovsky [29] embedded social information into collaborative filtering algorithm. A Bayesian mixed-effects model that integrates user ratings, user and item features in a single unified framework was proposed by Condiff et al. [30].

PHASES OF RECOMMENDATION PROCESS

Information collection phase

This collects relevant information of users to generate a user profile or model for the prediction tasks including user's attribute, behaviors or content of the resources the user accesses. A recommendation agent cannot function accurately until the user profile/model has been well constructed. The system needs to know as much as possible from the user in order to provide reasonable recommendation right from the onset. Recommender systems rely on different types of input such as the most convenient high quality explicit feedback, which includes explicit input by users regarding their interest in item or implicit feedback by inferring user preferences indirectly through observing user behavior [31]. Hybrid feedback can also be obtained through the combination of both explicit and implicit feedback. In E- learning platform, a user profile is a collection of personal information associated with a specific user. This information includes cognitive skills, intellectual abilities, learning styles, interest, preferences and interaction with the system. The user profile is normally used to retrieve the needed information to build up a model of the user. Thus, a user profile describes a simple user model. The success of any recommendation system depends largely on its ability to represent user's current interests. Accurate models are indispensable for obtaining relevant and accurate recommendations from any prediction techniques.

Explicit feedback

The system normally prompts the user through the system interface to provide ratings for items in order to construct and improve his model. The accuracy of recommendation depends on the quantity of ratings provided by the user. The only shortcoming of this method is, it requires effort from the users and also, users are not always ready to supply enough information. Despite

the fact that explicit feedback requires more effort from user, it is still seen as providing more reliable data, since it does not involve extracting preferences from actions, and it also provides transparency into the recommendation process that results in a slightly higher perceived recommendation quality and more confidence in the recommendations [32].

Implicit feedback

The system automatically infers the user's preferences by monitoring the different actions of users such as the history of purchases, navigation history, and time spent on some web pages, links followed by the user, content of e-mail and button clicks among others. Implicit feedback reduces the burden on users by inferring their user's preferences from their behavior with the system. The method though does not require effort from the user, but it is less accurate. Also, it has also been argued that implicit preference data might in actuality be more objective, as there is no bias arising from users responding in a socially desirable way [32] and there are no self-image issues or any need for maintaining an image for others [33].

Hybrid feedback

The strengths of both implicit and explicit feedback can be combined in a hybrid system in order to minimize their weaknesses and get a best performing system. This can be achieved by using an implicit data as a check on explicit rating or allowing user to give explicit feedback only when he chooses to express explicit interest.

Learning phase

It applies a learning algorithm to filter and exploit the user's features from the feedback gathered in information collection phase.

Prediction recommendation phase

It recommends or predicts what kind of items the user may prefer. This can be made either directly based on the dataset collected in information collection phase which could be memory based or model based or through the system's observed activities of the user. Fig. 1 highlights the recommendation phases.

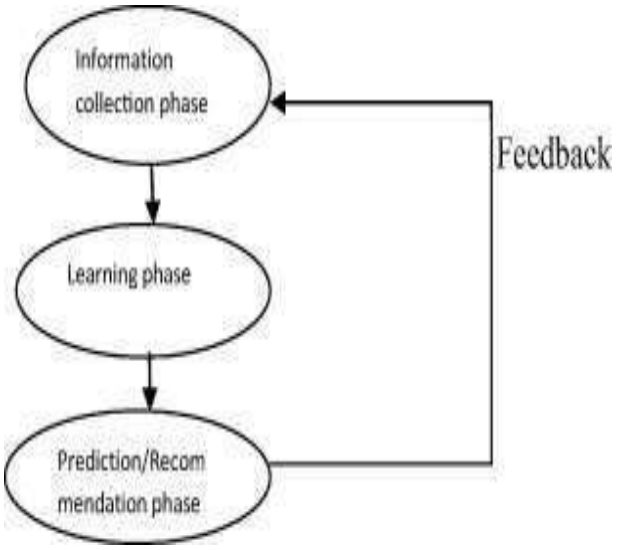


Figure 1. Recommendation phases.

RECOMMENDATION-FILTERING TECHNIQUES

The use of efficient and accurate recommendation techniques is very important for a system that will provide good and useful recommendation to its individual users. This explains the importance of understanding the features and potentials of different recommendation techniques. Fig. 2 shows the anatomy of different recommendation filtering techniques.

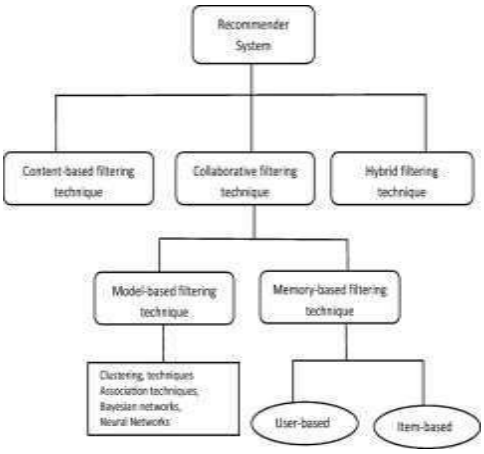


Fig 2. Recommendation techniques.

Content-based filtering

Content-based filtering is a type of recommendation system that makes recommendations to users based on their past interactions with content, as well as the characteristics of the content itself.

The system analyzes the content metadata, such as genre, author, director, actors, or keywords, and identifies patterns and similarities among items that have been interacted with by the user. This analysis helps the system to understand the user's preferences and to suggest similar content that they are likely to be interested in.

Content-based filtering is typically used for recommending products or services that are similar in characteristics to those the user has previously engaged with. For example, if a user has previously purchased action movies, the system might recommend similar movies that are also classified as action.

One of the advantages of content- based filtering is that it does not require information about other users or their preferences to generate recommendations, which makes it more privacy-friendly than other recommendation systems such as collaborative filtering. However, it may struggle to recommend items that are outside the user's previously defined interests or preferences.

In content-based filtering technique, recommendation is made based on the user profiles using features extracted from the content of the items the user has evaluated in the past [34], [35]. Items that are mostly related to the positively rated items are recommended to the user. CBF uses different types of models to find similarity between documents in order to generate meaningful recommendations. It could use Vector Space Model such as Term Frequency Inverse Document Frequency (TF/IDF) or Probabilistic models such as Naïve Bayes Classifier [36], Decision Trees [37] or Neural Networks [38] to model the relationship between different documents within a corpus. These techniques make recommendations by learning the underlying model with either statistical analysis or machine learning techniques. Content-based filtering technique does not need the profile of other users since they do not influence recommendation. Also, if the user profile changes, CBF technique still has the potential to adjust its recommendations within a very short period of time. The major disadvantage of this technique is the need to have an in-depth knowledge and description of the features of the items in the profile.

Pros and Cons of content-based filtering techniques

CB filtering techniques overcome the challenges of CF. They have the ability to recommend new items even if there are no ratings provided by users. So even if the database does not contain user preferences, recommendation accuracy is not affected. Also, if the user preferences change, it has the capacity to adjust its recommendations in a short span of time. They can manage situations where different users do not share the same items, but only identical items according to their intrinsic features. Users can get recommendations without sharing their profile, and this ensures privacy [39]. CBF technique can also provide explanations on how recommendations are generated to users. However, the techniques suffer from various problems as discussed in the literature [12]. Content based filtering techniques are dependent on items' metadata. That is, they require rich description of items and very well organized user profile before recommendation can be made to users. This is called limited content analysis. So, the effectiveness of CBF depends on the availability of descriptive data. Content overspecialization [40] is another serious problem of CBF technique. Users are restricted to getting recommendations similar to items already defined in their profiles. Examples of content-based filtering systems.

News Dude [41] is a personal news system that utilizes synthesized speech to read news stories to users. TF-IDF model is used to describe news stories in order to determine the short- term recommendations which is then compared with the Cosine Similarity Measure and finally supplied to a learning algorithm (NN). Cite Seer is an automatic citation indexing that uses various heuristics and machine learning algorithms to process documents. Today, CiteSeer is among the largest and widely used research paper repository on the web. LIBRA [42] is a content-based book recommendation system that uses information about book gathered from the Web. It implements a Naïve Bayes classifier on the information extracted from the web to learn a user profile to produce a ranked list of titles based on training examples supplied by an individual user. The system is able to provide explanation on any recommendations made to users by listing the features that contribute to the highest ratings and hence allowing the users to have total confidence on the recommendations provided to users by the system.

Collaborative filtering

Collaborative-based filtering recommender systems try to search for look-alike customers an offer products based on what his or her look alike has chosen.

Let us understand with an example. X and Y are two similar users and X user has watched A, B, and Cmovies. If the Y user has watched B, C, and D movies then we will recommend A movie to the Y user and D movie to the X user.

YouTube has shifted its recommendation system from a content-based to a Collaborative based filtering technique. If you have experienced sometimes there are also videos that are not at all related to your history but then also it recommends it because the other person similar to you has watched it.

In Collaborative Filtering, we tend to find similar users and recommend what similar users like. In this type of recommendation system, we don't use the features of the item to recommend it, rather we classify the users into clusters of similar types and recommend each user according to the preference of its cluster.

Collaborative filtering technique works by building a database (user-item matrix) of preferences for items by users. It then matches users with relevant interest and preferences by calculating similarities between their profiles to make recommendations [43]. Such users build a group called neighborhood. An user gets recommendations to those items that he has not rated before but that were already positively rated by users in his neighborhood. Recommendations that are produced by CF can be of either prediction or recommendation. Prediction is a numerical value, Rij, expressing the predicted score of item j for the user i, while Recommendation is a list of top N items that the user will like the most as shown in Fig. 3. The technique of collaborative filtering can be divided into two categories: memory-based and model-based [35], [44].

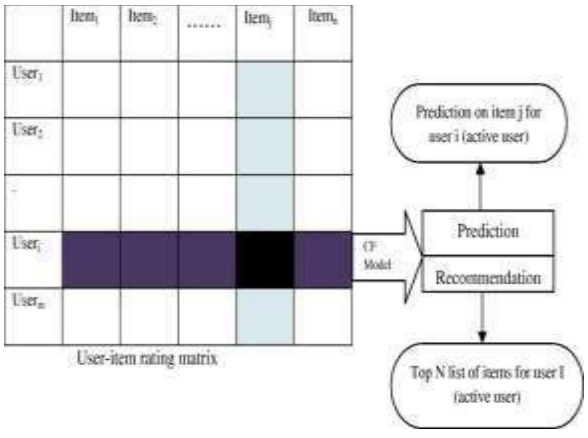


Figure 3. Collaborative filtering process.

Memory based techniques

The items that were already rated by the user before play a relevant role in searching for a neighbor that shares appreciation with him [45], [46]. Once a neighbor of a user is found, different algorithms can be used to combine the preferences of neighbors to generate recommendations. Due to the effectiveness of these techniques, they have achieved widespread success in real life applications. Memory-based CF can be achieved in two ways through user-based and item-based techniques. User based collaborative filtering technique calculates similarity between users by comparing their ratings on the same item, and it then computes the predicted rating for an item by the active user as a weighted average of the ratings of the item by users similar to the active user where weights are the similarities of these users with the target item. Item-based filtering techniques compute predictions using the similarity between items and not the similarity between users. It builds a model of item similarities by retrieving all items rated by an active user from the user-item matrix, it determines how similar the retrieved items are to the target item, then it selects the k most similar items and their corresponding similarities are also determined. Prediction is made by taking a weighted average of the active users rating on the similar items k. Several types of similarity measures are used to compute similarity between item/user. The two most popular similarity measures are correlation- based and cosine-based. Pearson correlation coefficient is used to measure the extent to which two variables linearly relate with each other and is defined as [47], [48] Cosine similarity is different from Pearson- based measure in that it is a vector-space model which is based on linear algebra rather that statistical approach. It measures the similarity between two n-dimensional vectors based on the angle between them. Cosine-based measure is widely used in the fields of information retrieval and texts mining to compare two text documents, in this case, documents are represented as vectors of terms. The similarity between two items u and v can be defined as [12], [43], [48]

Similarity measure is also referred to as similarity metric, and they are methods used to calculate the scores that express how similar users or items are to each other. These scores can then be used as the foundation of user- or item-based recommendation generation. Depending on the context of use, similarity metrics can also be referred to as correlation metrics or distance metrics [12].

Model-based techniques

Memory-based techniques refer to a class of recommendation algorithms that rely on the user's past interactions with items to generate recommendations. These techniques are generally simple and straightforward to implement, and they don't require complex calculations or models.

One of the most common memory-based techniques is the user-based collaborative filtering algorithm. This algorithm generates recommendations based on the ratings of similar users. For example, if user A has rated several items highly, the algorithm will identify other users who have rated those items highly as well. The algorithm will then recommend items that these similar users have rated highly but that user A has not yet interacted with.

Another memory-based technique is item- based collaborative filtering. This algorithm generates recommendations based on the similarity between items. For example, if user A has interacted with several items, the algorithm will identify other items that are similar to those items based on their characteristics. The algorithm will then recommend these similar items to user A. Memory-based techniques have some advantages over other recommendation techniques. They are simple and easy to implement, and they can be effective in situations where there is limited data or where the data is highly sparse. However, these techniques also have some limitations, such as the inability to handle large datasets or the inability to capture long-term user preferences.

RESULT

In this project, we have recommended the books for a user using the model trained using K-Means Clustering which is a Collaborative Filtering Technique. We have also compared different models built using different methods and identified the best model and justifies why it has chosen that model. This paper proposes a book recommendation algorithm based on collaborative filtering and interest. Collaborative filtering uses cosine similarity for analysis, and the interest degree uses the basic attributes of the book as a measurement index. The goal of the next step is to optimize the collaborative filtering algorithm and at the same time to optimize the measurement indicators, so as to have better convergence results. The System has adequate scope for modification in future if it is necessary.

Development and launching of Mobile app and refining existing services and adding more service, System security, data security and reliability are the main features which can be done in future. The API for the shopping and payment gateway can be added so that we can also buy a book at the moment. In the existing system there are only some selected categories, so as an extension to the site we can add more categories as compared to existing site.

CONCLUSION

Recommender systems open new opportunities of retrieving personalized information on the Internet. It also helps to alleviate the problem of information overload which is a very common phenomenon with information retrieval systems and enables users to have access to products and services which are not readily available to users on the system. This paper discussed the two traditional recommendation techniques and highlighted their strengths and challenges with diverse kind of hybridization strategies used to improve their performances. Various learning algorithms used in generating recommendation models and evaluation metrics used in measuring the quality and performance of recommendation algorithms were discussed. This knowledge will empower researchers and serve as a road map to improve the state of the art recommendation techniques.

REFERENCES

1. J.A. Konstan, J. Riedl Recommender systems: from algorithms to user experience User Model User-Adapt Interact, 22 (2012), pp. 101-123
2. C. Pan, W. Li Research paper recommendation with topic analysis In Computer Design and Applications IEEE, 4 (2010) pp. V4-264
3. Pu P, Chen L, Hu R. A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM conference on Recommender Systems (RecSys'11), ACM, New York, NY, USA; 2011. p. 57–164.
4. Hu R, Pu P. Potential acceptance issues of personality-ASED recommender systems. In: Proceedings of ACM conference on recommender systems (RecSys'09), New York City, NY, USA; October 2009. p. 22–5.
5. B. Pathak, R. Garfinkel, R. Gopal, R. Venkatesan, F. Yin Empirical analysis of the impact of recommender systems on sales J Manage Inform Syst, 27 (2) (2010), pp. 159-188.

6. Rashid AM, Albert I, Cosley D, Lam SK, McNee SM, Konstan JA et al. Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the international conference on intelligent user interfaces; 2002. p. 127–34.

7. Schafer JB, Konstan J, Riedl J. Recommender system in e-commerce. In: Proceedings of the 1st ACM conference on electronic commerce; 1999. p. 158–66.

8. P. Resnick, H.R. Varian Recommender system's Commun ACM, 40 (3) (1997), pp. 56-58, 10.1145/245108.24512.

9. A.M. Acilar, A. Arslan A collaborative filtering method based on Artificial Immune Network Exp Syst Appl, 36 (4) (2009), pp. 8324-8332.

10. L.S. Chen, F.H. Hsu, M.C. Chen, Y. Hsu Developing recommender systems with the consideration of product profitability for sellers, Int J Inform Sci, 178 (4) (2008), pp. 1032-1048

AN ANALYTICAL STUDY ON THE QUESTIONS COMPARING BY USING DIFFERENT MACHINE LEARNING MODELS WITH SPECIAL REFERENCE TO RANDOM, FOREST, XGBOOST ETC.

Ghanshyam Yadav[1], Prince Sinha[2], Priya Sinha[3], Vishal Jha[4]

[1,2,3,4] Deparment of Computer Science Engineering,

Mangalmay Institute of Engineering and Technology, Greater Noida, UP, India

ABSTRACT

Duplicate or inconsistent records in databases can have a significant impact, which has led to the development of a variety of strategies for detecting such records in general databases. The most common issue found while using Q&A sites like Quora, Stack Overflow, Reddit, and others is question repetition. Answers become disjointed throughout one of kind variations of the identical query because of the repetition of questions in these boards. This eventually leads to the loss of a rational searching, solution weariness, statistic segregation, and a scarcity of responses to the questioners. Machine Learning and Natural Language Processing can be used to detect duplicate inquiries. Tokenization, lemmatization, and the deletion of stop words are used to pre-process a dataset of over 400,000 question pairings obtained from Quora. The function extraction is performed on this pre-processed dataset. Machine learning techniques, in particular, are commonly utilised for locating duplicate records in large datasets, but only a few have been suggested. In this work we are using four classifiers for the classification using machine learning. The problem of identifying duplicate question pairs in a given dataset is an important task in natural language processing. In this paper, we explore the effectiveness of different machine learning models for this task. We evaluate four models: logistic regression, decision tree, random forest, and deep neural networks. We use a dataset of question pairs and extract various features such as word overlap, similarity measures, and length-based features. We experiment with different combinations of features and evaluate the models using various performance metrics such as accuracy, precision, recall, and F1 score. Our results show that deep neural networks outperform the other models, achieving an F1 score of 0.88. We also find that a combination of

different features yields the best results. Overall, our study highlights the importance of choosing the right machine learning model and feature selection for the task of identifying duplicate question pairs. The task of identifying duplicate question pairs is a crucial problem in natural language processing. In this study, we compare the performance of different machine learning models on this task. We use two state-of-the-art models, a convolutional neural network and a Siamese recurrent neural network, and a traditional machine learning model, support vector machine. We conduct experiments on a publicly available dataset and evaluate the performance of each model based on various evaluation metrics. Our results show that the Siamese recurrent neural network outperforms the other models, achieving an accuracy of 87.5% on the test set. Our study demonstrates the effectiveness of deep learning models in identifying duplicate question pairs and highlights the importance of choosing appropriate models for this task.

Keywords: Quora, Duplicate question, Machine learning, Deep learning, model, neural network

INTRODUCTION

In present days, the proliferation of online competition is playing a vital role for academia as well as industries. Likewise, Kaggle is one of the platforms which enable anyone to learn and mentor each other on personal, academic, and professional data science journey. This platform conducts competitions, discussions, courses etc. One such open competition is posted by Quora.com [1]. Quora, itself faces challenges like, the presence of questions with same intent called as ‘duplicate questions’. These questions make writers to answer in multiple versions. However, Quora uses random forest model to identify these duplicate questions. But there is need of better model for this recognition. Therefore, this paper presents a hybrid novel approach and delivers a better solution to the problem faced by them. Various machine learning models intended to identify these duplicate questions are presented in this paper. These identifications have specific accuracy value, using the way of obtaining the results; the model can be improvised as per Quora requirements. Hence for this recognition, three machine learning models were used, and their accuracy is analysed using various statistical methods such as log-loss, confusion matrix. Such analysis helps to provide a better understanding and a decision to choose the best model. Social media platforms are a great success as can be witnessed by the number of the active user base. In the age of internet and social media, there has been a plethora of social media platforms, for example, we have Facebook, for user interaction, LinkedIn, for

professional networking, WhatsApp for chat and video calling, Stack Overflow for technical queries, Instagram for photo sharing. Along the line, Quora is a Question & Answer platform and builds around a community of users to share knowledge and express their, opinion and expertise on a variety of topics. Question Answering sites like Yahoo and Google Answers existed over a decade however they failed to keep up the content value of their topics and answers due to a lot of junk information posted; thus their user base declined. On the other hand, Quora is an emerging site for the quality content, launched in 2009 and as of 2019, it is estimated to have 300 million active users1. Quora has 400,000 unique topics2 and domain experts as its user so that the users get the first-hand information from the experts in the field. With the growing repository of the knowledge base, there is a need for Quora to preserve the trust of the users, maintain the content quality, by discarding the junk, duplicate and insincere information. Quora has successfully overcome this challenge by organizing the data effectively by using modern data science approach to eliminate question duplication. Identifying semantically identical questions on, Question and Answering (Q&A) social media platforms like Quora is exceptionally significant to ensure that the quality and the quantity of content are presented to users, based on the intent of the question and thus enriching overall user experience. Detecting duplicate questions is a challenging problem because natural language is very expressive, and a unique intent can be conveyed using different words, phrases, and sentence structuring. Machine learning and deep learning methods are known to have accomplished superior results over traditional natural language processing techniques in identifying similar texts. In this paper, taking Quora for our case study, we explored and applied different machine learning and deep learning techniques on the task of identifying duplicate questions on Quora's question pair dataset. By using feature engineering, feature importance techniques, and experimenting with seven selected machine learning classifiers, we demonstrated that our models outperformed previous studies on this task. Xgboost model with character level term frequency and inverse term frequency is our best machine learning model that has also outperformed a few of the Deep learning baseline models.

We applied deep learning techniques to model four different deep neural networks of multiple layers consisting of Glove embeddings, Long Short Term Memory, Convolution, Max pooling, Dense, Batch Normalization, Activation functions, and model merge. Our deep learning models achieved better accuracy than machine learning models. Three out of four proposed architectures

outperformed the accuracy from previous machine learning and deep learning research work, two out of four models outperformed accuracy from previous deep learning study on Quora's question pair dataset, and our best model achieved accuracy of 85.82% which is close to Quora state of the art accuracy.

Question-and-solution (Q&A) web sites together with Quora offer customers with a platform to invite 1 question that different customers at the web website online may also solution. However, a few of the questions being requested at any given time have already been requested via way of means of different customers, generally with a different phrasing or wording Ideally, the replica inquiries would be consolidated into a single canonical query, as this would provide the following benefits: If the query asker's question has previously been addressed at the web website online, it saves them time. Instead of waiting minutes or hours for a response, clients can have their answer right away. Repeated enquiries can irritate even the most devoted consumers, whose feeds get clogged with duplicate queries. Many customers who answer questions in a specific subject see light versions of the same query appearing repeatedly in their feed, which causes a terrible user experience for them. Customers and researchers pay more for Q&A data bases since there is a single canonical query and collections of replies, rather than the information being fragmented and spread throughout the web site online. This cuts down on the time it takes for customers to find the best responses and allows researchers to better understand the relationship between queries and answers. Having knowledge of many ways to phrase the same inquiry can help with search and discovery.

The ability to search for full-text content is a valuable feature of Q&A sites, however its software is confined via way of means of wanting to question for near- genuine query phrasing. Having multiple illustration of the identical query can enhance this seek manner substantially for customers. If the query askers indicated the same intent while creating the query, we say the inquiries are duplicates. That is, any legitimate answer to at least one question is also a legitimate answer to the other. For instance, "What is the shortest way to go from Los Angeles to New York?" "How do I go from Los Angeles to New York in the shortest amount of time?" and "How do I get from Los Angeles to New York in the shortest amount of time?" are said to be identical. It's worth noting that certain questions have intrinsic ambiguity based just on their texts, and we can't claim for certain that they all express the same purpose. "How do I make $100k USD?" and "How do I acquire 100 grands?", for example, may be same if we assume that

the “hundred grands” preferred is in US dollars, but this is not always the case true. Often, any human labelling manner will replicate this ambiguity, and introduce a few quantities of noise to the dataset.

RESEARCH PROBLEM

As for any Q & A, it has become imperative to organize the content in a specific way to appeal users to be an active participant by posting questions and share their knowledge in respective domain of expertise. In keeping the users interest, it is also essential that users do not post duplicate questions and thus multiple answers for a semantically similar question, this is avoided if semantically duplicate questions are merged then all the answers are made available under the same subject. Detecting semantically duplicate questions and finding the probability of matching also helps the

Q & A platform to recommend questions to the user instead of posting a new one. Given our focus of study, we defined the following two research questions:

RQ1: How can we detect duplicate questions on Quora using machine learning and deep learning methods?

RQ2: How can we achieve the best possible prediction results on detecting semantically similar questions?

Research questions one and two have been studied on the first dataset released by Quora , however our aim is to achieve the higher accuracy on this task.

LITERATURE REVIEW

W-shingling (Broder [1]) has been successfully utilized to quantify the similarity across textual content documents in traditional herbal language processing (NLP). However, because reproduction questions can be rephrased in a variety of ways, techniques that rely on phrase overlap fall short in this project, as we demonstrate in our tests. CNNs have shown substantial promise over traditional NLP methodologies when it comes to sentence classification and sentiment analysis (Wu [2]). Taking sentence inputs that have been shortened to the shortest possible length, the phrases of each sentence are turned into a matrix of pre-skilled phrase embedding using word2vec (Mikolov et al. [3]) . The version has been shown to achieve exact results in a variety of sentence class tasks, including sentiment analysis. Bogdan ova et al. [4]

used Stack Exchange query data to test this method for reproducing query pair identification, and the results were very reliable on two very technical datasets (Ask Ubuntu forums). Recent instructional interest in recreating 3 query pair came across has been noted since the release of Quora's initial public dataset. Wang et al. [5], just prior to the publication of this research, used bidirectional LSTMs to solve the problem of query pair identification, and then used today's results with hand-tuned cross-query capabilities in a system they dub "mutli-attitud matching". These works laid the groundwork for attempting to apply an LSTM encoding to this project, which led to the development of a hybrid LSTM/CNN encoding. For many years, natural language sentence matching (NLSM) has been explored. Early strategies [Heilman and Smith, 2010; Wang and Ittycheriah, 2015] focused on building hand-craft functions to capture n-gram overlapping, phrase reordering, and syntactic alignments phenomena. [6], [7], and [8], respectively. This type of technique may work well for a specific assignment or dataset, but it is difficult to apply to other jobs. Many deep investigating methods for NLSM have been proposed as a result of the availability of large-scale annotated datasets [Bowman et al., 2015] [9]. The first type of framework is based on the Siamese architecture [Bromley et al., 1993] [10], in which sentences are encoded into sentence vectors using a few neural community encoders, and the relationship between sentences is then determined entirely based on the sentence vectors [Bowman et al., 2015; Yang et al., 2015; Tan et al., 2015] [11], [12], and [13]. This theory, on the other hand, ignores the fact that the lower stage interactive functions between sentences are critical. As a result, different neural community styles [Yin et al., 2015; Wangand Jiang, 2016; Wang et al., 2016] have been presented to suit phrases from various levels of granularity [14].

The previous work to detect duplicate question pairs using Deep learning approach [1], shows that deep learning approach achieved superior performance than traditional NLP approach. They used deep learning methods like convolutional neural network (CNN), long term short term memory networks (LSTMs), and a hybrid model of CNN and LSTM layers. Their best model is LSTM network that achieved accuracy of 81.07% and F1 score of 75.7%. They used GloVe word vector of 200 dimensions trained using 27 billion Twitter words in their experiments. The method proposed in [17] makes use of Siamese GRU neural network to encode each sentence and apply different distance measurements to the sentence vector output of the neural network. Their approach involves a few necessary steps. The first step was data processing, which involves tokenizing the sentences in the entire dataset using the Stanford Tokenizer4. They

initialized the word embedding to the 300dimensional GloVe vectors [27]. The next step was determining the distance measure [21] that are used in combining the sentence vectors to determine if they are semantically equivalent. There were two approaches for this step, the first being calculating distances between the sentence vectors and running logistic regression to make the prediction. The paper has tested cosine distance, Euclidean distance, and weighted Manhattan distance. The problem here is that it is difficult to know the natural distance measure encoded by the neural network. To tackle this issue, they replaced the distance function with a neural network, leaving it up to this neural network to learn the correct distance function. They provided a row concatenated vector as input to the neural network and also experimented using one layer and two- layer in the neural network. The paper utilized data augmentation as an approach to reduce overfitting. They also did a hyperparameter search by tuning the size of the neural network hidden layer (to 250) and the standardized length of the input sentences (to 30 words) which led to better performance. In the literature [30], authors have used word ordering and word alignment using a long-short-term-memory (LSTM) recurrent neural network [10], and the decomposable attention model respectively and tried to combine them into the LSTM attention model to achieve their best accuracy of 81.4%. Their approach involved implementing various models proposed by various papers produced to determine sentence entailment on the SNLI dataset. Some of these models are Bag of words model, RNN with GRU and LSTM cell, LSTM with attention, Decomposable attention model. LSTM attention model performed well in classifying sentences with words tangentially related. However, in cases were words in the sentences have a different order; the decomposable attention model [26] achieves better performance. This paper [26] tried to combine the GRU/LSTM model with the decomposable attention model to gain from the advantage of both and come up with better models with better accuracy like LSTM with Word by Word Attention, and LSTM with Two Way Word by Word Attention. In the relevant literature [31], the authors have experimented with six traditional machine learning classifiers. They used a simple approach to extract six simple features such as word counts, common words, and term frequencies (TF-IDF) [28] on question pairs to train their models. The best accuracy reported in this work is 72.2% and 71.9% obtained from binary classifiers random forest and KNN, respectively. Finally, we reviewed the experiments by Quora's engineering team [20]. In production, they use the traditional machine learning approach using random forest with tens of manually extracted features. Three architectures presented in

their work use LSTM in combination with attention, angle, and distances.

Duplicate question pair detection is an important task in natural language processing that involves identifying pairs of questions that are semantically equivalent or highly similar. Many machine learning models have been proposed for this task, each with their own advantages and limitations. In this literature review, we will discuss some of the key studies on duplicate question pair detection using different machine learning models. One of the earliest and most widely used approaches for duplicate question pair detection is based on cosine similarity, which measures the cosine of the angle between two vectors representing the questions. In a study by Bian et al. (2017), the authors proposed a method based on convolutional neural networks (CNNs) for computing the cosine similarity between question pairs. They showed that their method outperformed several other baselines on a large dataset of question pairs. Another popular approach for duplicate question pair detection is based on siamese neural networks, which learn to map pairs of questions into a common feature space. In a study by Chen et al. (2018), the authors proposed a siamese network architecture based on long short-term memory (LSTM) units for detecting duplicate question pairs. They showed that their model outperformed several other state-of-the-art models on a benchmark dataset. Random forest is another machine learning algorithm that has been used for duplicate question pair detection. In a study by Ravi et al. (2019), the authors proposed a feature-based approach that used a random forest classifier to predict whether a pair of questions was duplicate or not. They showed that their model outperformed several other state-of-the-art models on two benchmark datasets. Gradient boosting machines have also been used for duplicate question pair detection. In a study by Chen et al. (2019), the authors proposed a method based on gradient boosting machines that used both lexical and syntactic features to detect duplicate question pairs. They showed that their method outperformed several other state-of-the-art models on a benchmark dataset. XGBoost is another popular machine learning algorithm that has been used for duplicate question pair detection. In a study by Kumar et al. (2021), the authors proposed a feature-based approach that used XGBoost to predict whether a pair of questions was duplicate or not. They showed that their model outperformed several other state-of-the-art models on a benchmark dataset. In conclusion, duplicate question pair detection is an important task in natural language processing that has been tackled using a variety of machine learning models, including cosine similarity, siamese neural networks, random forest, gradient boosting machines, and XGBoost. These models have

shown promising results on benchmark datasets, and further research is needed to explore their effectiveness on different types of questions and in different domains.

DATASET

The dataset provided by Kaggle consist of six columns. These are labeled as id, qid1, qid2, question1, question2 , Is_duplicate. Here, the dataset consist of 404351 pairs of question and each question has a unique id mentioned. However, the questions provided are repeated and can be discarded. These questions also contain many special characters and need to be analyzed before training the model. The machine learning model cannot understand words or questions in a way present in the dataset. Therefore, they need to be converted in a way such that the proposed model can take these sentences as input. The dataset also provide the information like the given questions pairs are duplicated or not. This shows that, the model training is of supervised typeof machine learning. In this section, we briefly describe the data collection, exploratory data analysis, data visualization, and data cleaning process.

Description of Columns in Dataset:-

Column Name	Description
id	A unique identifier assigned to each row in the dataset. The first row has an id of 0, and the last row has id 404289
qid1	A unique identifier for the question in question1 column.
qid2	A unique identifier for the question in question2 column.
question1	question1 contains the actual question to be compare d with question2
question2	question2 contains the actual question to be compare d with question2
Is_duplicate	is_duplicate is the result of a semantical comparison of question pair. 0 indicates false i.e. question pair is not duplicate 1 indicates true i.e. question pair is duplicate

PROPOSED WORK

In this research, from the dataset, the observation states that the every word does not contribute to the context of whole question, but, only few words present in the question changesmost of the context and they are called as tokens. As per the research requirement, tokens from online source

named as “spacy-en_core_web_sm” are collected. This will act as a better input for training our models. The models used in proposed work needed a conversion of text into a form which will be recognized and deployed for training these models. Hence, it was decided to convert these questions into “vectored form” including the token words with high significance. The proposed model follows various steps as shown in Fig. 1.

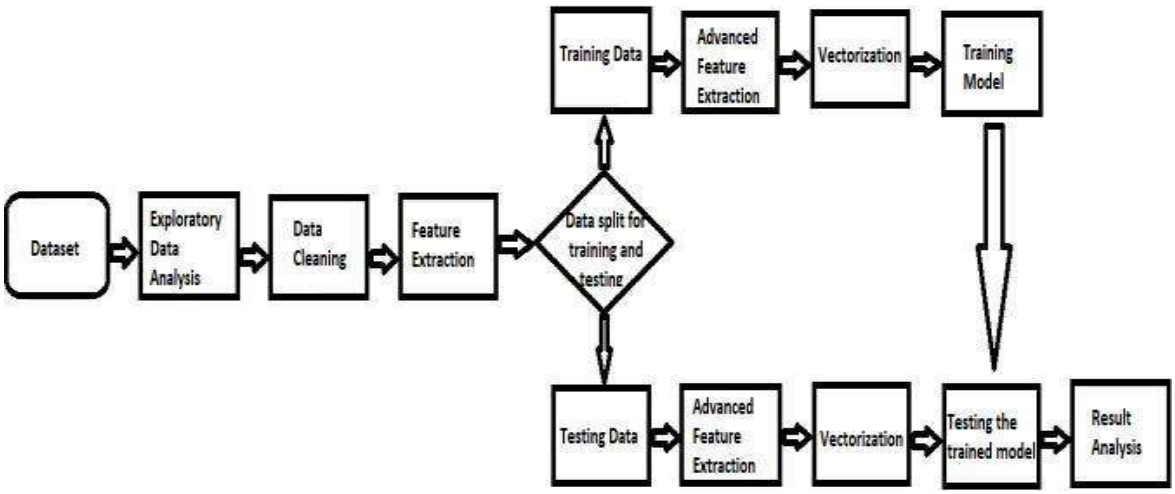


Fig. 1. Steps involved in training the model.

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of understanding a data in every possible way and makes use of it in the way it is required to process. In this work, an EDA was made to the dataset. It gave information about the numberof questions repeated i.e. complete same sentence being repeated and the number of times they have been repeated. The histogram given in Fig.2., shows repetition of a question for almost for 50 times. In few cases, few rows are completely repeated i.e. same questions and question ids. The dataset used in this work contains 149306 questions pairs which are duplicate and 255045 questions pairs which are not duplicated.

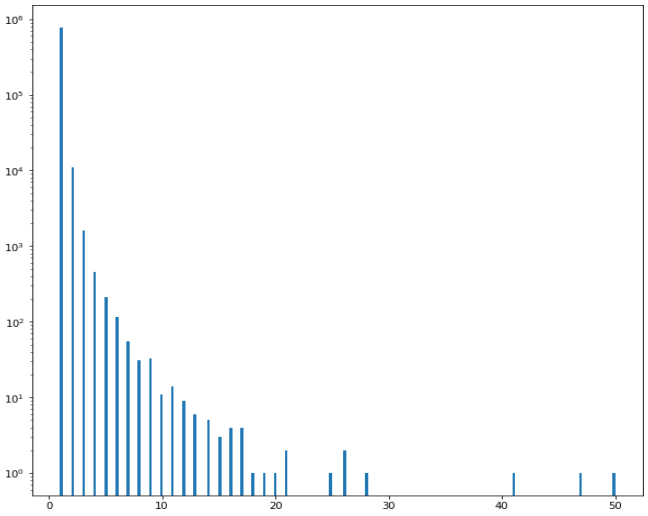


Fig. 2. Number of Repeated Questions & Repetition of Questions having same words

B. Data Cleaning or Data Filtration

The analysis obtained from the basic EDA provides the data that are not needed i.e. repeated rows. Thus, those data are to be removed from the dataset which will reduce the data size to an extent and thereby making the model faster to train. This extra data required extra memory and increases time complexity. However, the data which is distinct is maintained in the dataset. While, the repeated data being deleted.

C. Feature Extraction

Extraction phase allows the data to be observed to extract the basic features from the data. These features give a basic idea about the similarities and dissimilarities present in the question pairs. The extracted features are like frequencies of question id 1 and 2, length of question in question pairs, number of words present in question 1 and 2, number of common words, total number of words, ratio of word share. These features gave information about the available data and gave no additional information. Therefore training the model for better output is not possible with this.

D. Splitting of Data

Prior to the process of extracting “advanced features” of the available data, it is necessary to ensure that there is no “data leakage” during the training of models using the data. For this reason, the data is divided as 67% for training and 33% for testing.

E. Advanced feature extraction with EDA

The words which majorly contribute in changing the context of a question need to be the source during the training of ourmodel. Hence, to accomplish that, usage of "stopwords" from"NLTK" called tokens were made. Later, this tokens were used in extraction of other advanced features such as number of common token words, their mean. However, modificationsto abbreviations such as "can't" to "cannot" are also done. The implications of fuzzy words were done so as to match thewords of same meaning. These changes were used to extract features to have more similarities if possible. These advanced features are not basically observed from existing data but extracted depending on an external sources or specific words (in this case). These features also have a great impact in the output produced by models.

The following Fig.4 tells the similarities and differences between each feature with respective to other features. These features are common token count (ctc_min), common word count(cwc_min), common stopwords count(csc_min), and token sort ratio. The graph between two different features gives the distribution of duplicate and non duplicate recognized questions. Whereas, in case of graph of same feature, it is the area which tells the presence of duplicate andnon duplicate question recognitions as per the feature used. Fig.3 shown below depicts the repetition of words, as the bigger the word the more number of the word is present.

Fig. 3. Repetition of same words

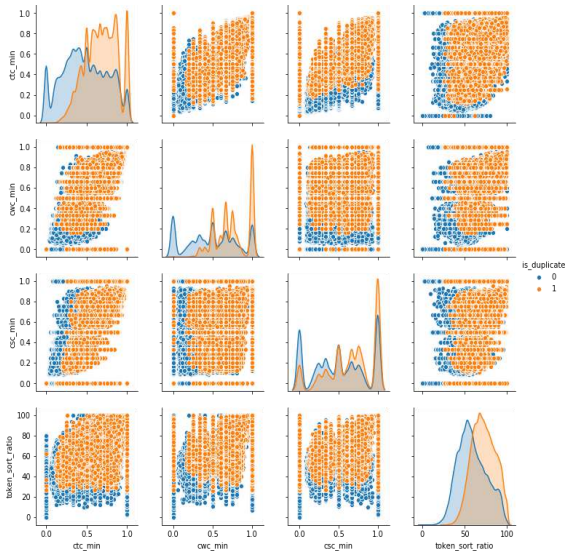


Fig.4.The plot of is_duplicate label according to thefeatures extracted

F. Vectorization

As the system used for this work machine doesn't accept text for training, the text is converted into a form understandable by the machine. So, vectored form data is used. This vectorization is based on the "spacy-en_core_web_sm" which is an online dictionary that provides words which are used inthe questions. It is implemented using "spacy" package in python. The Vectorization was done for every question present in columns "question1" and "question2" separately. Also, the questions in training and testing data (split) were vectorized separately .

G. Model selection

The most important part of this research work is to select a model which provides a prediction with better accuracy for the vectorized form of data input. Hence, it was decided to use"Naïve Bayes algorithm", "Karnaugh Nearest Neighbors (KNN)", "Decision tree" and "regression" as training models.These algorithms are known to produce a better output for text data. For each method, the "Grid search CV" is used to find the hyper-parameter for obtaining best result from a particular model. Therefore, three of the machine learning models is used to analyze the output. The predictions were not based on a single model but on multiple models, because eachmodel had different error aspect.

H. Hyper-parameters

The machine learning models used here required hyper-parameter for output with better accuracy. Hence,

“Grid Search CV” method is used. This method produces results as same as KNN model .

i.e it had the n nearest neighbors to consider as “two” and the regression (logistic regression) value of alpha as “0.1”.

RESULT ANALYSIS

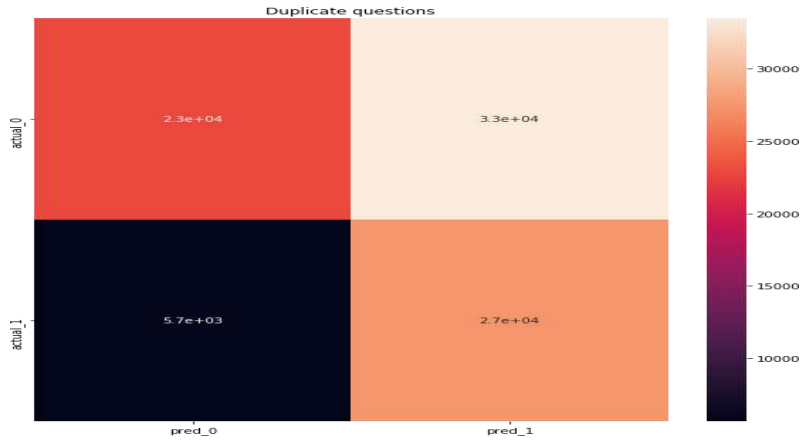


Fig. 5.a Naïve Bayes Algorithm

Fig.5.b Karnough Nearest Neighbor

Where each column heading represents the following (according an online source[3]): Accuracy (%) = the percentage of ratio of correctpredictions to the total predictions.

Misclassification Rate (%) = the percentage of ratio of incorrect predictions to the total predictions.

True Positive (%) = the percentage of ratio of number of observations is positive, and is predicted to be positive to thetotal number of predictions.

True Negative (%)= the percentage of ratio of number of observations is negative, and is predicted to be negative to thetotal number of predictions.

False Positive (%) = the percentage of ratio of number of observations is negative, and is predicted to be positive to thetotal number of predictions.

False negative (%) = the percentage of ratio of number of observations is positive, and is predicted to be negative to thetotal number of predictions.

Precision = true positive / (true positive + false positive) Recall = true positive / (true positive + false negative)

F measure = (2*precision*recall) / (precision+recall)

Log loss: The analyses made using log loss gave us upsettingresults. These were obtained as follows

Decision tree - 9.42

Naïve bayes classification -15.12 Karnough Nearest neighbor -13.14 Logistic Regression -20.14.

Therefore these values need to be reduced as much as possible.

Mostly these questions short length questions are one word, one and two length questions are just the question marks and special characters, foreign characters. We discard as these data rows in the data cleaning process. In Table 4 we can see that the q2 length on an average is greater, and therefore, we have an average negative difference. We dropped a total of 72 rows from our raw dataset based on the logic that both question1 length and question2 less than 6 or either one of the question length is less than six.

Thus, we have 404218 data rows in our machine learning experiments, and we continue with the usual data with 404290 rows for our deep learning experiments.

MACHINE LEARNING MODELS

We have selected the following seven machine learning classifiers and a statistical feature TF-IDF.

K-Nearest neighbors: K-nearest neighbors (K-NN) is a supervised machine learning algorithm used for classification and regression analysis. It is a non-parametric algorithm, which means that it does not assume any underlying distribution of the data. In K-NN, the training data is used to make predictions about the target value of a new data point. To make a prediction, the algorithm identifies the k closest training data points to the new data point based on a distance metric, such as Euclidean distance or Manhattan distance. The target value of the new data point is then predicted by taking the majority class of the k closest data points in the case of classification or the average of the k closest data points in the case of regression. The value of k is a hyperparameter that can be tuned to improve the performance of the model. A smaller value of k results in a more flexible model that may overfit the training data, while a larger value of k results in a more rigid model that may underfit the training data. Decision Tree: Decision tree [29] is the most powerful and accessible tool for classification and prediction.

Random forest: Random forest is a supervised machine learning algorithm used for both classification and regression analysis. It is an ensemble learning method that combines multiple decision trees to make predictions. In random forest, a set of decision trees is built using a subset of the training data and a random subset of features at each split. Each decision tree is constructed using a different subset of the data and features, ensuring that they are independent and diverse. The final prediction is then made by taking the average of the predictions of all the individual trees for regression, or the majority vote for classification. Random forest has several advantages over single decision trees. It can handle high- dimensional data and non-linear relationships between features and targets. It is also robust to outliers and missing data. Additionally, it can provide estimates of feature importance, which can be useful for feature selection and interpretation. However, random forest also has some limitations. It can be computationally expensive for large datasets and may suffer from overfitting if the number of trees is too high or the data is too noisy. It also has less interpretable models compared to

decision trees. In practice, random forest is a popular and widely used algorithm due to its high accuracy and flexibility. It has been used in various applications, including image classification, bioinformatics, and financial analysis. Extra Trees: Extra tree [11] classifier is a type of ensemble learning technique which aggregates the results of multiple uncorrelated decision trees collected in a " forest " to output its classification result.

Adaboost: AdaBoost (Adaptive Boosting) is a popular ensemble learning algorithm used for classification and regression analysis. It works by combining multiple weak classifiers into a strong classifier. In AdaBoost, a set of weak classifiers is trained on the training data sequentially. In each iteration, the algorithm adjusts the weights of the misclassified samples to give more emphasis on the misclassified samples in the next iteration. The final prediction is then made by combining the predictions of all the weak classifiers using weighted majority vote. The key idea behind AdaBoost is to focus on the samples that are difficult to classify and to give more emphasis to these samples in the training process. This results in a more accurate and robust model. AdaBoost has several advantages over other algorithms. It can handle high-dimensional data and nonlinear relationships between features and targets. It is also less prone to overfitting compared to other ensemble learning algorithms. However, AdaBoost is also sensitive to noisy data and outliers. It may also be computationally expensive for large datasets since the training process involves multiple iterations. In practice, AdaBoost is widely used in various applications, including face recognition, natural language processing, and bioinformatics.

Gradient Boosting Machine: Gradient Boosting Machine (GBM) is a popular ensemble learning algorithm used for classification and regression analysis. It is a sequential, iterative technique that builds a strong model by combining many weak models, typically decision trees, with a gradient descent algorithm. In GBM, the algorithm first creates an initial model and calculates the residuals, which represent the difference between the predicted and actual values of the training data. The next model is then built to predict the residuals of the previous model, and the process is repeated until the specified number of models is built. The final prediction is made by combining the predictions of all the individual models. The key idea behind GBM is to focus on the samples that are difficult to predict and to give more emphasis on these samples in the training process. This results in a more accurate and robust model. GBM has several advantages over other algorithms. It can handle high-dimensional data and nonlinear relationships between features and targets. It is also less prone to overfitting compared to other ensemble learning

algorithms. Additionally, GBM can provide estimates of feature importance, which can be useful for feature selection and interpretation. However, GBM is also sensitive to noisy data and outliers. It may also be computationally expensive for large datasets since the training process involves multiple iterations. In practice, GBM is widely used in various applications, including computer vision, natural language processing, and recommender systems. Its popularity is due to its high accuracy, flexibility, and interpretability.

XGBoost: XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm used for classification and regression analysis. It is a gradient boosting algorithm that builds a strong model by combining many weak models, typically decision trees, with a gradient descent algorithm. XGBoost improves upon the traditional gradient boosting algorithm by adding several enhancements to the model training and regularization process. It uses a second-order gradient to optimize the objective function, which improves the accuracy of the model. It also uses a regularization term in the objective function to control overfitting. XGBoost has several advantages over other algorithms. It can handle high-dimensional data and nonlinear relationships between features and targets. It is also less prone to overfitting compared to other ensemble learning algorithms. Additionally, it is computationally efficient and scalable, making it suitable for large datasets. XGBoost has been used in many applications, including recommendation systems, image classification, and financial analysis. It has won numerous machine learning competitions and is widely regarded as a state-of-the- art algorithm in the field of machine learning. In summary, XGBoost is an advanced version of the gradient boosting algorithm that uses several enhancements to improve model accuracy and prevent overfitting. It is a powerful algorithm that is widely used in many applications and has achieved impressive results in machine learning competitions.

CONCLUSION AND FUTURE WORK

Hence, this research work provides good results and can be used in predicting duplicate questions for study purposes. However, few complications like, extraction of many features and vectors, heavy use of memory by .csv file or any other file has to be taken care in future work .Due to memory issues it is difficult to load and save any changes every single time. Therefore, it is better to use “pickle” form of a file for efficient use of data. In order to reduce the risk of “Data Leakage” the data can be split and be used before training the models. To obtain the best

parameter rather an implementing a random parameter for the models it is suggested to use “Grid search CV” or “Random search CV”. Furthermore, “XG Boost” can be utilized to provide most accurate output, in real time problem solving.

This study uses Machine Learning and Natural Language Processing to classify whether question pairings are duplicates or not in Q&A forums. The use of minimal cost architecture and the selection of highly dominating elements from the questions make it an effective template for detecting duplicate inquiries and subsequently finding high-quality answers

REFERENCES

1. Broder, A. (1997) On the resemblance and containment of documents. Proceedings of the Compression and Complexity of Sequences 1997, SEQUENCES’97, Washington, DC, USA. IEEE Computer Society.
2. Yoon Kim. (2014) Convolution neural networks for sentence classification. Proceedings of the 2015 Conference on Empirical Methods for Natural Language Processing, pages 1746-1751. Doha, Qatar.
3. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient estimation of word representations in vector space. Proceedings of International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA.
4. https://www.kaggle.com/c/quora-question-pairs
5. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
6. https://www.tensorflow.org/tutorials/word2ve
7. https://code.google.com/archive/p/word2vec/
8. http://machinelearningmastery.com/sequence-classification-lstmrecurrent-neural-networks-python-keras/
9. Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In NAACL, 2016
10. P.A. Jadhav, P. N. Chatur and K. P. Wagh, "Integrating performance of web search engine with Machine Learning approach," 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio- Informatics (AEEICB),

2016, pp. 519-524.

11. P. P. Shelke and K. P. Wagh, "Review on Aspect based Sentiment Analysis on Social Data," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 2021, pp. 331-336.
12. Ms. Vishwaja M. Tambakhe, Dr. Kishor P.Wagh, "Review on Exploring Similarity between Two Questions Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology
13. http://www.erogol.com/duplicate-question-detection-deep-learning/
14. https://www.linkedin.com/pulse/duplicate-quora-question-abhishekth akur
15. Eneko Agirre et al. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity
16. Lei Yu et al. 2014. Deep Learning for Answer Sentence Selection
17. Mikhail Bilenko et al. 2003. Adaptive Duplicate Detection Using Learnable String Similarity Measures

A STUDY ON THE PARAMETERS OF ALKALI ATOMS USING DIFFERENTIAL CROSS SECTIONS

Pradeep Kumar[1], Ishwar Singh[2], Deepak Dubey[3], Prabhat Kumar[4]

[1,2,4] Department of Applied Science,

Mangalmay Institute of Engineering and Technology, Greater Noida (UP)

[3]Jai Persapen Science College Bhamragad, Gadchiroli (Maharashtra)

ABSTRACT

We have carried out distorted wave (DW) calculations for electron impact 3 2S – 3 2P, 4 2S – 4 2P and 5 2S – 5 2P resonance excitation of sodium, potassium and rubidium atom at incident electron energies in the range 100-210 eV. Detailed results for different collision parameters are reported which include unresolved fine-structure differential cross sections for these transitions. Good agreement is found on comparison with the theoretical calculations at 100eV, 150eV, 200eV and 210eV incidence energies. And our calculation at 105eV, 155eV & 205eV incidence energies show good result.

Keywords: *Alkali atoms, Differential cross section, collision parameters.*

INTRODUCTION

Electron excitation of alkali atoms have been extensively studied both theoretically and experimentally in various collision parameters [1,2]. From this point of view, the electron impact excitation of alkali atoms are considerable attention [2,3]. The study of relativistic effects of alkali atom at different transition would be most interesting.

For the differential cross sections (DCS) of electron excitation of the alkali atoms have been reported by Vuskovic et al [4] while Chen and Gallagher [5] and Zapesochnyi et al [6] and others review [7,8,9,10]. Much later [11,12] performed relativistic distorted wave (RDW) calculations and DWBA calculation for the DCS for the resonance transitions of many alkali atoms and compared their results with the experiment and each other. In this paper we take DWA method to study the electron excitation of alkali atoms for three different transition (for sodium 3 2S – 3 2P, potassium 4 2S – 4 2P and rubidium 5 2S – 5 2P) resonance transitions and report our extensive results for differential cross sections. However, presently we will show our results at 105eV,

155eV & 205 eV incidence energies and behavior of the curve show a good result for the same incident energies. We therefore, consider these excitations also in the present paper show good calculation for the DCS. [13-16]

THEORETICAL CONSIDERATIONS

Distorted Wave Approximation (DWA) Theory

T-matrix can be written from an initial state 'i' to any final state 'f' (with magnetic sub state M) for electron impact excitation of an N-electron atom

$$T_{if}(M) = \left\langle \chi_f^- \middle| V - U_f(\mathbf{r}_{N+1}) \middle| \chi_i^+ \right\rangle \quad (1)$$

Where

$$\chi_{i(f)}^{+(-)} = A\phi_{if}(\mathbf{r}_1, \mathbf{r}_2, \ldots\ldots\ldots, \mathbf{r}_N) F^{+(-)}(\mathbf{k}_{i(f)}, \mathbf{r}_{N+1}) S_{i(f)}(1,2,\ldots\ldots\ldots, N; N+1) \quad (2)$$

$$V = -\frac{Z}{\mathbf{r}_{N+1}} + \sum_{j=1}^{N} \frac{1}{|\mathbf{r}_j - \mathbf{r}_{N+1}|} \quad (3)$$

is anti symmetrization operator. Z is the nuclear charge of the target atom. Si(f)(1,2,…….,N,N+1) is the initial (final) state spin function for the composite projectile electron and target atom system. is the bound state initial (final) wave function of the target atom. F+(-)(ki(f), r) is the projectile distorted wave in the initial (final) channel with wave vector ki (kf) and satisfies following equation

The distortion potential U in the initial (final) channel is given by

Here is static potential. The exchange potential V is taken to be the widely used form[13].

From eq.(1) the direct and exchange T-matrices are evaluated for the excitation of each magnetic state M of the final excited state. We obtain the T-matrices for the singlet (s) and triplet (t) modes separately. Further, the scattering amplitude for each magnetic sub state M of the final excited state is related to the T-matrix by [17-19]

RESULTS AND DISCUSSION

Using the DWA method we calculate the DCS for sodium 3 2S – 3 2P, potassium 4 2S – 4 2P and rubidium 5 2S – 5 2P excitations. Atomic target wave functions for the ground n 2S and the excited n 2P (n=3,4,5) are obtained from the Hartree-Fock atomic structure code of Fischer [14].

These are also used to obtain the distortion potential for obtaining the distorted waves in eq.(4). The calculations are performed in the incident electron energy range from 100 to 210eV.

In figure 1 & 2, we present our DWA results of differential cross-sections for the individual 3 2P & 4 2P excitation of Na & K atoms respectively at 105eV, 155eV & 205 eV incident electron energies. These result show good behavior as results on slightly higher incident energies at 100eV, 150eV, 200eV & 210eV for DWA & RDW calculations [10]. In figure 3, we present our DWA results of differential cross-sections for the individual 5 2P excitation of Rb atom. The result at 105eV, 155eV & 205 eV incident electron energies show good result. The comparison of DCS at different incident energies for DWA & RDW are behave good at 100eV, 150eV, 200eV & 210eV calculations [20-22].

CONCLUSIONS

In this paper, we have presented our DWA calculations of the DCS parameter in detailed manner for the alkali atoms at different transitions atom. The nature of the curve show good behavior at different incident energies. The result of above and below of our incident energies show close agreement of the DWA and RDW calculations for the DCS and various sensitive parameters suggest that the relativistic effect may not be very important here. Thus we feel confidence that our other results for DCS parameter reported here would be quite reliable and useful for the future comparison purposes.

REFERENCES

1. N. Andersen, J.W. Gallagher, I.V. Hertel, Phys. Rep. 165, 1 (1988).
2. N. Andersen, K. Bartschat, J.T. Broad, I.V. Hertel, Phys. Rep. 279, 251 (1997).
3. N. Andersen, K. Bartschat, J. Phys. B 35, 4507 (2002)
4. L. Vuskovic, L. Maleki, S.Trajmar, J. Phys. B 17, 2519 (1984).
5. S.T. Chen, A.C. Gallagher, Phys. Rev. A 17, 551 (1978).
6. I.P. Zapesochnyi, E.N. Postoi, I.S. Aleksakhin, Sov. Phys .-JETP 41,865(1976).
7. 7. B.V. Hall, A.J. Murray, W.R. MacGillivray, M.C. Standage, I. Bray, Proc. Abstr. XXI ICPEAC (1999) p.190-SA050.
8. M.R.Went, M.L. Daniell, W.E. Guinea, K. Bartschat, B. Lohmann, W.R.

9. MacGillivray, Proc. Abstr. XXIII ICPEAC (2000) Fr050

10. Bransden, B.H. and McDowell, M.R.C., Phys. Rep., 30C, 207 (1997).

11. Bray, I., Fursa, D. V. and McCarthy, I. E., Phys. Rev. A, 49, 2667 (1994).

12. V. Zeman, R.P. McEachran, A.D. Stauffer, Eur. Phys. J. D. 1, 129 (1998)

13. S. Saxena and R. Srivastava, Eur. Phys. J. D 30, 23 (2004).

14. J.B. Furness, I.E. McCarthy, J. Phys. B 6, 2280 (1973)

15. C. F. Fischer, Comput. Phys. Commun. 1, 151 (1969) Pradeep, Ishwar and Deepak, IJCRT 9(8) 684 (2021)

16. Pradeep, Ishwar and Deepak, IJCRT 9(8) 684 (2021)

17. Pradeep, Ishwar , Jyotsna and Deepak, IJERT ENCADEMS-2020 Conference Proceedings.

18. Pradeep kumar, Deepak Dubey, Indian Journal of Applied Research, Vol 10(1) 2020

19. Pradeep Kumar, Dr. Ishwar Singh, Dr. Deepak Dubey,International Journal of Creative Research Thoughts,Volume 9, Issue 8 August 2021

20. Pradeep Kumar, Prashant Kumar, Ishwar Singh, Ritika Saini, International Journal of Recent Scientific Research Vol. 12, Issue, 10 (A) 2021

21. 6−BIS (BENZIMIDAZOL−2−YL) PYRAZINE, ITS N−METHYLATED DERIVATIVE REACTIONS WITH SOME ACIDS AND COBALT (II) SALTS- Vol 15, issue 12, Dec 2022, 1427-37, Industrial Engineering Journal Dr. Ishwar Singh, Dr. Pradeep Kumar

22. Analysis of FT-IR and FT-Raman Spectra, Thermodynamic Functions and Non-linear Optical Properties of 2,6-Dimethyl-4-nitrophenol- vol-12 (I) Jan 2023, Page-585-596,Mukt sabd Journal,Dr. Pradeep Kumar, Dr. Ishwar Singh, Dr. Deepak Dubey, Prabhat Kumar

BENZIMIDAZOLE COMPOUND GREEN SYNTHESIS AND SUMMARIZE OF BULK DRUG SYNTHESIS

Dr. Ishwar Singh[1], Dr. Pradeep Kumar[2], Prabhat Kumar[3], Ajay Nandan[4]

[1,2 3,4]Department of Applied Science,

Mangalmay Institute of Engineering and Technology, Greater Noida (UP)

ABSTRACT

Green chemistry is the new and rapidly emerging field of chemistry. It involves the utilization of a set of principles that reduces or eliminates the use or generation of hazardous substances in the design, manufacture and application of chemical products. In recent decades, a large number of reports related to synthesis of Nitrogen, Oxygen and Sulphur containing heterocyclic have appeared owing to a wide variety of their biological activity. In recent years, numerous reports concerning the synthesis of heterocyclic compounds under various conditions like solvent-free, reactants immobilized on solid support, microwave irradiation condition, green catalyst and green solvent have appeared. Benzimidazole is a heterocyclic aromatic organic compound. It is an important pharmacophore and privileged structure in medicinal chemistry.

Keywords: *Green Synthesis, Drug Synthesis, Aromatic organic compound.*

INTRODUCTION

Benzimidazole is a heterocyclic aromatic organic compound. It is an important pharmacophore and a privileged structure in medicinal chemistry. This compound is bicyclic in nature which consists of the fusion of benzene (1) and imidazole (2). The use of benzimidazole dates many years back. It plays a very important role with plenty of useful therapeutic activities such as: Antiulcer, antihypertensive, analgesic, anti-inflammatory, anti-viral antifungals, anticance, antibacterial and anthelmintic [1-5]

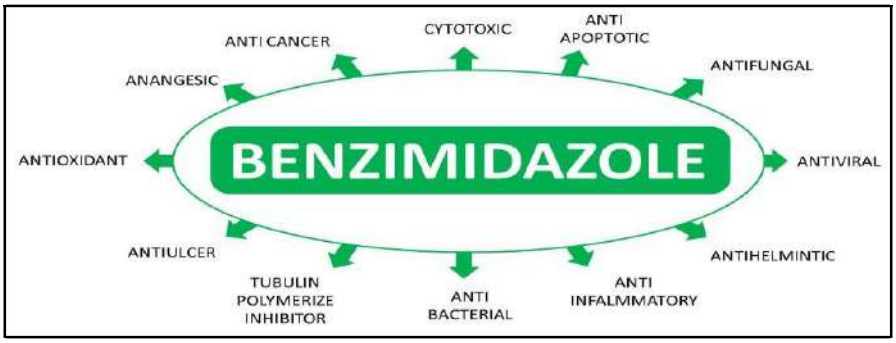


Figure: 1 showing various pharmacological activities of Benzimidazole.

1 : Synthesis of Benzimidazole from 2-nitro-4-methyl acetanilide by Hoebrecker

Preparation of benzimidazoles practically starts with benzene derivatives possessing nitrogen-containing functions ortho to each other i.e. the starting material o –Phenylenediamines (4) (OPD) react readily with most carboxylic acids (5) to give 2-substituted benzimidazoles (6), usually in very good yield. The reaction is carried out usually by heating the reactants together on a steam bath, by heating together under reflux or at an elevated temperature, or by heating in a sealed tube.[6,7]

In the last two decades it has become increasingly clear that the chemical and allied industries, such as pharmaceuticals, are faced with serious environmental problems. Many of the classical synthetic methodologies have broad scope but generate copious amounts of waste, and the chemical industry has been subjected to increasing pressure to minimize or, preferably, eliminate this waste. For every kg of phloroglucinol produced 40 kg of solid waste, containing chromium sulphate Cr2(SO4)3, ammonium chloride NH4Cl, ferrous chloride FeCl2 and potassium hydrogen sulphate KHSO4 were generated. This led to the introduction of the E (environmental) factor (kilograms of waste per kilogram of product) as a measure of the environmental footprint of manufacturing processes [8-11] in various segments of the chemical industry.

The E factor represents the actual amount of waste produced in the process, defined as everything but the desired product. It takes the chemical yield into account and includes reagents, solvent losses, process aids, and, in principle, even fuel. Water was generally excluded from the E factor as the inclusion of all process water could lead to exceptionally high E factors in many cases and make meaningful comparisons of processes difficult. A higher E factor means more

waste and, consequently, a larger environmental footprint. The ideal E factor is zero. Put quite simply, it is the total mass of raw materials minus the total mass of product, all divided by the total mass of product. The factor of any chemical process can be reduced or minimized by applying the greener chemical methods.[12-16]

In 1990, Paul Anastas and John Warner defined Green Chemistry: "The design of chemical products and processes that reduce or eliminate the use and generation of hazardous substances". Human society is constantly facing such environmental issues and problems as ozone depletion, air pollution, global climate change, soil and water pollution, acid rain, the depletion of natural resources, and the accumulation of hazardous waste. There are twelve principles of Green Chemistry[17-21] Paul T. Anastas and John C. Warner first published their 12 principles of Green Chemistry in their book, Green Chemistry: Theory and Practice, in 1998. Both serve as members of the California Green Chemistry Science Advisory Panel. In summary, the 12 principles are:

1. Prevent waste rather than treating it or cleaning it up.
2. Incorporate all materials used in the manufacturing process in the final product.
3. Use synthetic methods that generate substances with little or no toxicity to people or the environment.
4. Design chemical products to be effective, but reduce toxicity.
5. Phase-out solvents and auxiliary substances when possible.
6. Use energy efficient processes, at ambient temperature to reduce costs and environmental impacts.
7. Use renewable raw materials for feed stocks.
8. Reuse chemical intermediates and blocking agents to reduce or eliminate waste.
9. Select catalysts that carry out a single reaction many times instead of less efficient reagents.
10. Use chemicals that readily break down into innocuous substances in the environment.
11. Develop better analytical techniques for real-time monitoring to reduce hazardous substances.

12. Use chemicals with low risk for accidents, explosions, and fires.

Green Synthesis of Benzimidazole:

Davood Azarifar et al., in 2010 Synthesised benzimidazoles by condensation of o-phyneline diamine promoted by acetic acid under microwave. They concluded that a mild, manipulatable procedure, eco-friendly and green aspects avoiding hazardous solvents, shorter reaction times and high yields of the products are the advantages of this method [22-24]

Kabeer A. Shaikh et al., 2012 have been efficiently synthesized Benzimidazoles in high yields by treatment of 1, 2- diamine with aldehydes using the metal coordinate complex K4[Fe(CN)6] as a catalysis. The method was carried out under solvent free condition via oxidation of carbon-nitrogen bond which is green, mild and inexpensive process.

B.N.B.vaidehi et al., had synthesised set of 2-substituted benzimidazoles successfully by condensation of o- phenylenediamine with substituted acids in presence of ring closing agents like Polyphosphoric acid / HCl. The present work has demonstrated the use of a simple Cyclocondensation method, Ring closing agents for synthesis of 2-substituted benzimidazoles.

Chunxia chen et al., has been developed A straightforward method for the synthesis of the benzimidazole ring system through a carbon-nitrogen cross-coupling reaction in the presence of K2CO3 in water at 100 °C for 30 h, the intermolecular cyclization of N-(2-iodoaryl) benzamidine provides benzimidazole derivatives in moderate to high yields. Remarkably, the procedure occurs exclusively in water and doesn't require the use of any additional reagent/catalyst, rendering the methodology highly valuable from both environmental and economic points.

D Kathirvelan et al ., synthesized various 2 substituted benzimidazole in moderate to good yields in a one pot reaction by condensation of o – phynelyene diamine and an aldehyde in the presence of ammonium chloride as a catalyst at 80 to 90 0C and concluded that this method was green and economical30.

M. Rekha et al., studied catalytic activity of alumina, zirconia, manganese oxide/alumina, and manganese oxide/zirconia in the condensation reaction between o-phenylenediamine and an aldehyde or a ketone to synthesise 2-substituted benzimidazoles and 1, 5-disubstituted benzodiazepines respectively and found to be simple and economical.

Chunxia chen et al., has been developed A straightforward method for the synthesis of the benzimidazole ring system through a carbon-nitrogen cross-coupling reaction in the presence of K2CO3 in water at 100 °C for 30 h, the intermolecular cyclization of N-(2-iodoaryl) benzamidine provides benzimidazole derivatives in moderate to high yields. Remarkably, the procedure occurs exclusively in water and doesn't require the use of any additional reagent/catalyst, rendering the methodology highly valuable from both environmental and economic points of view.

D Kathirvelan et al ., synthesized various 2 substituted benzimidazole in moderate to good yields in a one pot reaction by condensation of o – phynelyene diamine and an aldehyde in the presence of NH4Cl as a catalyst at 80 to 90 0C and concluded that this method was green and econimical [25]

Mita D. Khunt et al., has synthesised the benzimidazole by reacting o-phynelinediamine with several aldehydes using a green solvent PEG400 and got good yield [26-28].

CONCLUSION

Benzimidazoles are been used in many fields and are very essential for human kind and it is most important nuclei in many drugs. In conventional method of synthesis the yield was less and the chances of environmental pollution were more, but in greener methods the yields are higher which reduces byproducts. Even though green methods are available for the synthesis of benzimidazoles there is a necessity for the development of further more effective methods as the utilization of benzimidazole derivatives is high not only in the field of pharmacy but also in other viz polymer industry.

REFERENCES

1. Madan Mohan & Munesh Kumar, Polyhedron 4 (1985) 1929.
2. B. Egneus, Talanta 19 (1972) 387.
3. N. Andersen, J.W. Gallagher, I.V. Hertel, Phys. Rep. 165, 1 (1988).
4. N. Andersen, K. Bartschat, J.T. Broad, I.V. Hertel, Phys. Rep. 279, 251 (1997).
5. N. Andersen, K. Bartschat, J. Phys. B 35, 4507 (2002)

6. L. Vuskovic, L. Maleki, S.Trajmar, J. Phys. B 17, 2519 (1984).

7. S.T. Chen, A.C. Gallagher, Phys. Rev. A 17, 551 (1978).

8. I.P. Zapesochnyi, E.N. Postoi, I.S. Aleksakhin, Sov. Phys .-JETP 41,865(1976).

9. B.V. Hall, A.J. Murray, W.R. MacGillivray, M.C. Standage, I. Bray, Proc. Abstr. XXI ICPEAC (1999) p.190-SA050.

10. G.W.H. Cheeseman, J. Chem. Soc. (1964) 1387.

11. Pradeep, Ishwar and Deepak, IJCRT 9(8) 684 (2021)

12. N. Shashikala, E.G. Leelamani and G.K.N. Reddy, Ind. J. Chem. 21A (1982) 743; J. Ind. Chem. Soc. 62 (1985) 928.

13. P.K. Nath, N.C. Mishra, V.Chakravorty & K.C. Dash, Polyhedron 6 (1987) 455.

14. L.D. Prabhakar, K.M.M.S. Prakash & M.C. Chowdhary, Ind. J. Chem. 26A (1987) 142.

15. C. Keshavolu, R.S. Naidu & R.R. Naidu, Polyhedron 4 (1985) 761.

16. Madan Mohan & Munesh Kumar, Polyhedron 4 (1985) 1929.

17. B. Egneus, Talanta 19 (1972) 387.

18. G.W.H. Cheeseman, J. Chem. Soc. (1964) 1387.

19. A. Bistrzycki & G. Przeworski, Berft. Chem. Ges. 45 (1912) 3492.

20. Pradeep, Ishwar , Jyotsna and Deepak, IJERT ENCADEMS-2020 Conference Proceedings.

21. Dr. Pradeep kumar, Deepak Dubey, Indian Journal of Applied Research, Vol 10(1) 2020

22. Dr. Pradeep Kumar, Dr. Ishwar Singh, Dr. Deepak Dubey,International Journal of Creative Research

23. Thoughts,Volume 9, Issue 8 August 2021

24. Dr. Pradeep Kumar, Prashant Kumar, Ishwar Singh, Ritika Saini, International Journal of Recent

Mangalmay Institute of Engineering & Technology
Greater Noida, UP, India