

# **AI/ML BASED LIBRARY**

**A Project Report**

Submitted in partial fulfillment of the  
requirements for the award of the Degree of

**MASTER OF SCIENCE (INFORMATION TECHNOLOGY)**

**By**

**RAMKRISHNA SARVANKAR**

Seat Number: 20306A1005

**Under the esteemed guidance of**

**Mrs. Beena Kapadia**

**Assistant Professor, Department of Information Technology**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**VIDYALANKAR SCHOOL OF INFORMATION TECHNOLOGY**

**(Affiliated to University of Mumbai)**

**MUMBAI, 400 037**

**MAHARASHTRA**

**2021 - 2022**

**VIDYALANKAR SCHOOL OF INFORMATION TECHNOLOGY**

**(Affiliated to University of Mumbai)**

**MUMBAI-MAHARASHTRA-400037**

**DEPARTMENT OF INFORMATION TECHNOLOGY**



**CERTIFICATE**

This is to certify that the project entitled, "**AI/ML BASED LIBRARY**", is bona fide work of **RAMKRISHNA SARVANKAR** bearing Seat No: 20306A1005 submitted in partial fulfilment of the requirements for the award of degree of MASTER OF SCIENCE in INFORMATION TECHNOLOGY from University of Mumbai.

**Internal Guide**

**Coordinator**

**Internal Examiner**

**External Examiner**

**Date:**

**College Seal**

**Principal**

University of Mumbai



University of Mumbai

16<sup>th</sup> Inter-Collegiate/  
Institute/Department



**AVISHKAR**

Research Convention: 2021-22

**SELECTION ROUND**

(District/Zonal Level Research Proposal Competition)

## *Certificate of Participation*

This is to certify that **Mr. Sarvankar Ramkrishna** of **Vidyalankar School of Information Technology, Wadala** has participated and submitted a research proposal titled **Artificial Intelligence / Machine Learning Based Smart Library — A Smart Way to Learn in Engineering and Technology** category and **PG level** for the **Selection Round of 16th Inter-Collegiate / Institute / Department Avishkar Research Convention: 2021-22** for zone.

  
**Dr. (Mrs.) Minakshi Gurav**  
OSD,  
Avishkar Research Convention,  
University of Mumbai

  
**Dr. Suhil Patil**  
Director,  
Department of Students' Development,  
University of Mumbai

Mumbai  
April 25, 2022



University of Mumbai



University of Mumbai

16<sup>th</sup> Inter-Collegiate/  
Institute/Department



Research Convention: 2021-22

**SELECTION ROUND**

(District/Zonal Level Research Proposal Competition)

## *Certificate of Merit*

This is to certify that **Mr. Sarvankar Ramkrishna** of **Vidyalankar School of Information Technology, Wadala** has participated and submitted a research proposal titled **Artificial Intelligence / Machine Learning Based Smart Library — A Smart Way to Learn in Engineering and Technology** category and **PG** level for the **Selection Round** of **16th Inter-Collegiate / Institute / Department Avishkar Research Convention: 2021-22** for **Mumbai-1** zone. The said research proposal is selected for **Final Round** of **16th Inter-Collegiate/Institute/Department Avishkar Research Convention: 2021-22**.

  
**Dr. (Mrs.) Minakshi Gurav**  
OSD,  
Avishkar Research Convention,  
University of Mumbai

  
**Dr. Suhil Patil**  
Director,  
Department of Students' Development,  
University of Mumbai



Mumbai  
April 25, 2022



University of Mumbai



University of Mumbai

16<sup>th</sup> Inter-Collegiate/  
Institute/Department



Research Convention: 2021-22


FINAL ROUND

(University Level Research Proposal Competition)

## *Certificate of Participation*

This is to certify that **Mr. Sarvankar Ramkrishna** of **Vidyalankar School of Information Technology, Wadala** has participated and presented a research proposal titled **Artificial Intelligence / Machine Learning Based Smart Library — A Smart Way to Learn in Engineering and Technology** category and **PG** level for the **Final Round** of **16th Inter-Collegiate / Institute / Department Avishkar Research Convention: 2021-22** organised on Virtual Platform by University of Mumbai on **May 2, 2022**.

  
**Dr. (Mrs.) Minakshi Gurav**  
OSD,  
Avishkar Research Convention,  
University of Mumbai

  
**Dr. Sunil Patil**  
Director,  
Department of Students' Development,  
University of Mumbai

May 2, 2022  
Mumbai





# Research Paper

## Study on Methodologies for Information Extraction

**Ravi Yadav**

Student,  
M Sc IT,

Vidyalankar School of Information Technology,  
Wadala(E), Mumbai  
Email: ravi.yadv@vsit.edu.in

**Ramkrishna Sarvankar**

Student,  
M Sc IT,

Vidyalankar School of Information Technology,  
Wadala(E), Mumbai  
Email: ramkrishna.sarvankar20@vsit.edu.in

**Guided By:**

**Mrs. Beena Kapadia**

Assistant Professor

Dept. of Information Technology

Vidyalankar School of Information Technology,  
Wadala(E), Mumbai

### ABSTRACT

The new era of Information has brought a lot of new opportunities in the domain of automation. As lots of new data in various formats are being stored and then mined for useful insights one such format which has created the deepest impact in overall automation domain is mining textual data to get the important stuff out of it and then use that for various analysis and decision making to various levels of automation. In this paper various methods for Information Extraction will be discussed and one of the methods in conclusion will be chosen for implementation.

**Keywords**— *Information Extraction, Data Mining, Automation, NLP, Information Processing.*

### I. INTRODUCTION

Information Extraction is one of the main domains of research in Artificial Intelligence. Various methodologies that use syntactical properties of the text combined with Machine Learning has been implemented. Two types of such methodologies that have been discussed here are Rule Based Information Extraction Methods and Classification based Extraction Methods. Rule based Information Extraction methods basically use syntactical rules to extract needful Information whereas Classification based Extraction methods try to implement different techniques to classify important and non-important data from the given text.

### II. LITERATURE REVIEW

NLP based Information Extraction methods have been there since 19's and various works have been performed as well. There are several successful methodologies that is being used today in various text related applications such as AutoSlog (Riloff, 1993) that uses a sentence parser to parse the natural language, Crystal (Soderland, 1995), (LP)2 (Ciravegna, 2001), etc.

### III. OBJECTIVE

Objective of doing this research paper was to find a method or a combination of them from, a whole library of methods, that can be implemented in our project which requires a method that is able to extract information from a given sentence using as less data as possible to get trained, with capabilities of domain specific learning and information extraction.

### IV. RESEARCH METHODOLOGY

This paper is a cumulative study of various methods based on NLP techniques to Extract Information which has been taken from various research papers published by the experts of the domain. Rigorous study has been performed to analyze the referenced papers and a comparative method has been used to identify the method that suits our project needs. This paper in a nutshell provides the comparative study of two types of NLP based Information Extraction Methods which are:

- 1) Rule Learning Based Extraction Method
- 2) Classification Based Extraction Method

There are other ways as well such as Sequential Based Extraction Method which are mostly for generalizing purpose which does not fulfil the domain specific requirement of our project thus being the reason of not being one of the models to study.

#### **Rule Learning Based Extraction Method:**

For Rule Learning Based Extraction Method, we went through the paper of one of the domain experts which is Information Extraction: Methodologies and Applications (Tang et al) which provides in detail explanation with subtopics and generalized idea of the Rule Learning Based Extraction Method. Below we provide a brief understanding of the topic:

In general, the methods can be grouped into three categories: dictionary-based method, rule-based method, and wrapper induction.

Rule Based System the rule-based method uses several general rules instead of dictionary to extract information from text. Two main rule learning algorithms of these systems are: bottom-up method which learns rules from special cases to general ones, and top-down method which learns rules from general cases to special ones. There are proposed many algorithms, such as (LP)<sup>2</sup> (Ciravegna, 2001), iASA (Tang, 2005b)

It learns two types of rules that respectively identify the start boundary and the end boundary of the text to be extracted. The learning is performed from the examples in a user-defined corpus (training data set). Training is performed in two steps: initially a set of tagging rules is learned; then additional rules are induced to correct mistakes and imprecision in extraction.

- **Dictionary Based Method:**

Traditional information extraction systems first construct a pattern (template) dictionary, and then use the dictionary to extract needed information from the new untagged text. These extraction systems are called as dictionary-based systems (also called pattern-based systems) including: AutoSlog and autoslog-ts.

The key point in the systems is how to learn the dictionary of patterns that can be used to identify the relevant information from a text. Each AutoSlog concept node has a conceptual anchor that activates it and a linguistic pattern, which, together with a set of enabling conditions, guarantees its applicability.

AutoSlog needs to parse the natural language sentence using a linguistic parser. The parser is used to generate syntax elements of a sentence (such as subject, verb, preposition phrase). Then the output syntax elements are matched against the linguistic pattern and fire the best matched pattern as the result pattern to construct a pattern dictionary.

**Classification Based Extraction Methods:**

The basic idea is to present information extraction problem as that of the classification.

Boundary detection using classification model

This system consists of two distinct phases:

- **Learning**

In the learning phase our system uses a set of labelled documents to generate models which we can use for future predictions.

- **Extracting**

The extraction phase takes the learned models and applies them to new unlabelled documents using the learned models to generate extractions.

This model aims to generate the boundaries (start boundary and end boundary) for the special information.

For IE to text the whole document is converted into tokens such as words or sentences. Then two classifiers are learned one for start and one for the end boundary in positive and negative fashion



where the start classifier marks positive the start boundaries and negative to the other whereas the second classifier does the vice-versa.

There have been many modifications to the above given methodology to overcome the problems related to it, which are discussed in the conclusion section below.

## CONCLUSION

This section of the paper discusses the selection of the final model for the project based on the pros and cons as per the project requirements of the models.

Rule Learning Based Model as described above is one of the models that can be used to Extract Information based on a dictionary that the model itself creates using a syntax parser. The benefit of such model is that if the extracted dictionary pattern is of good differentiators, then the model will work best, but if the dictionary gets affected then the whole model will have to suffer. Building a good syntax parser hence becomes much more important than Extracting Information.

(Tang-et-al)Classification Based Extraction Method as discussed above is an extraction method which classifies information with other data using boundary classification. This model works well when the dataset is small but if data becomes large then the probability of the identified boundary to be right becomes very low, which shows that to make this model work we will require a model with high precision otherwise we would be stuck with just false positives. There are solutions to this problem such as, use of two-level boundary and others.

As classification-based methods were aiming towards a more general domain of natural language processing, which basically becomes much bigger than our project domain. On the other hand, Rule Learning Based Extraction is what seems feasible seeing the projects needs of being domain specific and low data training capacity. So, we decided to go with the Rule Learning based Extraction keeping our window open for bootstrapping in future as well.

## References

1. Tang-et-al - Information Extraction: Methodologies Applications
2. AutoSlog (Riloff, 1993)

# ACKNOWLEDGEMENT

It gives me immense pleasure in expressing my heartfelt thanks to the people who were part of this project in numerous ways. I owe my thanks to all those who gave endless support right from the conception of the project idea to its implementation, it would not have materialized without the help of many.

The dedication, hard work, patience, and correct guidance makes any task proficient & a successful achievement. Intellectual and timely guidance not only helps in trying productive but also transforms the whole process of learning and implementing into an enjoyable experience.

I would like to thank our principal “**Dr. Rohini Kelkar**” and vice principal “**Mr. Asif Rampurawala**” for providing this opportunity, a special thanks to our MSc IT coordinator “**Ms. Beena Kapadia**” for their support, blessings and for being a constant source of inspiration to us. With immense gratitude, I would like to convey my special honor and respect to “**Ms. Beena Kapadia**” who took keen interest in checking the minute details of the project work and guided us throughout the same.

A sincere thanks to the non-teaching staff for providing us with the long lab timings that we could receive along with the books and with all the information we needed for this project, without which the successful completion of this project would not have been possible.

Finally, I wish to avail this opportunity & express a sense of gratitude and love to my friends and my beloved parents for their support, strength and help for everything.

**Mr. RAMKRISHNA SARVANKAR**

# DECLARATION

I hereby declare that the project entitled, **AI/ML BASED LIBRARY** done at Vidyalankar School of Information Technology, has not been in any case duplicated to submit to any other universities for the award of any degree. To the best of my knowledge other than me, no one has submitted to any other university.

The project is done in partial fulfillment of the requirements for the award of degree of **MASTER OF SCIENCE (INFORMATION TECHNOLOGY)** to be submitted as final semester project as part of our curriculum.

Name and Signature of the Student

# Table of Contents

<b>CHAPTER 1: INTRODUCTION.....</b>	<b>1</b>
1.1. Background .....	2
1.2. Objectives .....	2
1.3. Purpose Scope & Applicability.....	3
1.3.1 Purpose.....	3
1.3.2 Scope.....	3
1.4. Applicability .....	3
• Technical Feasibility .....	3
• Operational Feasibility.....	4
<b>CHAPTER 2: LITERATURE REVIEW.....</b>	<b>5</b>
2.1. Introduction.....	6
2.2. Literature Review.....	6
2.3. Comparative Analysis .....	7
2.4. Research Gap .....	7
<b>CHAPTER 3: SYSTEM DESIGN &amp; MODELLING .....</b>	<b>8</b>
3.1. Introduction.....	9
3.2. Information Requirement Gathering and Analysis .....	9
3.3. System Modelling .....	10
3.4. Database analysis and modelling .....	12
3.5. System diagram.....	13
<b>3.5.1. Basic Module .....</b>	<b>13</b>
<b>3.5.2. Use Case Diagram .....</b>	<b>13</b>
<b>CHAPTER 4: METHODOLOGY.....</b>	<b>14</b>
4.1. Introduction .....	15
4.2. Block diagram/Flowchart of proposed methodology .....	15
4.3. Description of methodology used.....	16
<b>CHAPTER 05: RESULTS, ANALYSIS AND DISCUSSION .....</b>	<b>18</b>
5.1. NLP Based Chatbot:.....	19
5.2. Video Search Feature: .....	20
5.3. Admin Side Module:.....	21
<b>CHAPTER 06: CONCLUSION, LIMITATIONS, AND FUTURE WORK .....</b>	<b>24</b>



5.1. Conclusion .....	25
5.2. Limitation.....	26
5.3. Future Work .....	26
<b>Summary .....</b>	<b>27</b>
<b>References.....</b>	<b>28</b>

## Table Of Figure

FIGURE 1 DATABASE MODEL.....	12
FIGURE 2 : BASIC MODULES.....	13
FIGURE 3 USE CASE DIAGRAM.....	13
FIGURE 4 :FLOWCHART FOR NLP BASED CHATBOT .....	15
FIGURE 5 FLOWCHART FOR VIDEO SEARCH .....	15
FIGURE 6: NLP BASED CHATBOT OUTPUT.....	19
FIGURE 7: VIDEO SEARCH MODULE OUTPUT .....	20
FIGURE 8:ADMIN SIDE MODULE OUTPUT .....	22

## List Of Tables

TABLE 1:EVENT TABLE.....	23
--------------------------	----



# **CHAPTER 1: INTRODUCTION**

## **1.1.Background**

Today in this digital era we have been making everything digital which has made many day-to-day tasks much easier and faster and library system has also been taken under its transformation. With all the digital media for study that we have such as pdf, eBooks, videos and many more. Now the whole digital system has been filled with these kinds of useful resources leading to great problem which is of finding the right resource for a particular need. So as Machine Learning and automation is making all the digital processes easy and fast done automatically, this same mentality has gone behind this project as well which will let a user find the right resources as per its needs fast and easy.

## **1.2.Objectives**

This project has been inspired by the very problem that students and teachers face while looking for a particular knowledge or resource and seeing the very impact of Machine learning and automation techniques that it is bringing in the whole digital industry. So, the objectives of this project are as follows

- 1) Bringing ease in searching process
- 2) Creating an interactive platform for resource finding

Providing different formats of resources in the same place with automation



### **1.3.Purpose Scope & Applicability.**

#### **1.3.1 Purpose**

There were many students who were facing issues in finding the right resource for a particular topic as per their level in the whole digital learning medium which was making them to waste their more time in searching for the right resource rather than in studying them. So, the purpose of this project is to make this searching process as much easy and interactive as possible using Machine learning and various automation techniques.

#### **1.3.2 Scope**

As there are not much useful ways to find the resources on the internet right now except some searching techniques which help people to find the resource but still requires a lot of checking and confirming, which shows a lot of scope for this project especially in the environment where everything is shifting towards digital medium.

### **1.4.Applicability**

- **Technical Feasibility.**

Machine Learning based Library as name suggest requires machine learning and some other techniques to implement the various automation that this project is trying to accomplish. By doing a cross checking with all the technological requirements and other requirements with all the different techniques available and which are implementable this project concludes to be technically feasible. The techniques that are required and with the ones that they are been implemented are as follows

Video search feature -> Machine Learning or Frame Matching

Chatbot -> NLP based techniques and other programming techniques

UI -> Visual basic UI framework or Python based UI framework

Database -> Either on premise MongoDB or Cloud Based Mongo DB

- **Operational Feasibility.**

The operational need of this project is basically a learned model which will help in searching the video and some NLP based techniques that will make the chatbot running. The project being a pc software and all the resources and processing of searching and responding to all the requests being done on cloud the user will not have to take the burden of various processing and other backend needs. With today's advance level of optimization and operational efficiency of the programs with multi fold increase in the processing power, our project concludes of being operationally feasible.

## **CHAPTER 2: LITERATURE REVIEW**

## **2.1.Introduction**

Machine Learning based Library project takes help from various methods and studies done by various domain experts, to make this project a reality. There is mixing and parameter tuning of these methods to make it according to the project needs.

## **2.2.Literature Review**

NLP based Information Extraction methods have been there since 19's and various works have been performed as well. There are several successful methodologies that are being used today in various text related applications such as autoslog (Riloff, 1993) that uses a sentence parser to parse the natural language, Crystal (Soderland, 1995), (LP<sup>2</sup>), (Ciravegna, 2001), etc.

All these and other papers on NLP techniques help us to understand and implement the technology appropriately for the NLP based chatbot that allows the user to find its resource interactively.

For video searching companies like Google and Facebook have been continuously working creating some or the other video related processing methods which help in navigating in the video for various searching applications. Techniques such as face matching using OpenCV and textual understanding in frame with translation using other machine learning based library has been implemented before that take full advantage of processing videos to give different applications.

So, for the other feature that is Video searching using screenshot various libraries of Python language based on image and video processing have been used with appropriate programming.



### **2.3.Comparative Analysis**

Comparing and analysing the methods and techniques implemented in this project with many of the previous work done on this or related topic gives various insights and results that helped us to improve on the methods implemented on this project by either implementing hybrid techniques or improving by providing home-made data. Some of the comparative analysis results were

- 1) Using domain specific data collected using survey and other methods to train the NLP model rather than using industry standard way of implementing model trained previously on general data.
- 2) Video searching using frame segregation rather than making a model learn to distinguish the frame.
- 3) Using only image processing rather than going with industry standard hybrid method of image processing with machine learning.

### **2.4.Research Gap**

As far as our project goes, comparing it with various papers and studies done online people and domain experts have been working on these kinds of methods with different applications have covered the various technological gap but the idea that we are implementing these techniques on generally shows a great research gap as there have been no instances where resource finding used these techniques to ease the whole process of searching the resources.

## **CHAPTER 3: SYSTEM DESIGN & MODELLING**

### **3.1.Introduction**

Designing a system that requires learning and has automation in its core values gets dependent on the data that it is modelled and build upon. So, gathering information for this project required generating information from very scratch by using methods such as survey and collaboratively designing questions for training the NLP model for Information Extraction as well as creating frames directly from video lectures to extract video directly from screenshot.

### **3.2.Information Requirement Gathering and Analysis**

Information Requirement gathering started with the requirement list that we had for the projects core features such as NLP based chatbot mainly for Information Extraction and the Video clip extraction from screenshot.

For building the NLP model the requirement was of information for all the relevant questions that we required to train the model which can extract Information relevant to provide the desired resources. This created a problem of creating the data for training the model from scratch as the data required was not available online or through any other resources. So, we went with surveying people for questions datasets, asking people for various types of queries they might be interested in to query to a system like this and what kind of answers they will be expecting so that if there would be a space for improvement in the system, we could make those for better user experience and satisfaction. We did our survey using both online and offline mediums using forms and verbal questioning We analysed the data that we acquired from survey and extracted all the relevant information, the analysis resulted into following findings, The various questions that we acquired resulted into a great dataset with various combinations of keywords and other sorts of details which made training the model much easier and better than initially thought training process.

For building the video search feature we required first the videos and then data of frames that will be using to search for the videos. So, we used a program that created frames from every the is being uploaded and then those frames are attached with a timestamp of the video. The overall data for video search feature was mainly the video frames and timestamps attached to them.

### 3.3.System Modelling

System Modelling included modelling different parts of the projects modelling separately as all the features of the model works on different modelling. The first modelling we did was of video search feature which included modelling the system in a way which can easily retrieve videos from the screenshot.

Phases of Modelling the Video Search feature:

- I. Choosing the methodology
- II. Creating the Data
- III. Building the model
- IV. Testing the model
- V. Evaluating the model

#### I. Choosing the Methodology:

There were many methods that were doing the sort of thing that we wanted to perform but not were as specific as we wanted to as per our feature, so we went with hybrid kind of solution where we build a model that actually creates frame of a video attaching its timestamp to it and then keeps them into the database where the search program searches the particular images from to search for video.

#### II. Creating the Data:

Data that we used to get the feature working was a data of video frames made from scratch from all the lecture videos that we could acquire for our database. To create this data of video frames a program written in python was used which after certain interval used get a copy of frame of the video and used to store it for searching purpose.

#### III. Building the Model:

The main modelling of search program that would search for the video lectures from the screenshots was modelled using python programming language which has various modules that were first creating the frames and then the main model uses to run on these pictures in future.

#### IV. Testing the model

NLP chatbot as it uses a more of a traditional method mixed with Machine Learning to extract information, we tested the NLP model using various techniques which mostly included testing techniques based on machine learning paradigms. Accuracy and precision were the judging scores for model evaluation.

For video search feature as the model is based on a matching algorithm the model is tested with human judgement and with techniques such as text matching results to understand how well the overall model works.

#### V. Evaluating the model

As evaluation is something that requires practical testing of the project with various people to test whether the application and its features are responding well with the various different ways in which people can ask the questions, and the results can also be checked if that is what the user expected, which requires the project to be completed at least 80 percent which will take some more time but the evaluation on paper of the project seems feasible.

### 3.4. Database analysis and modelling

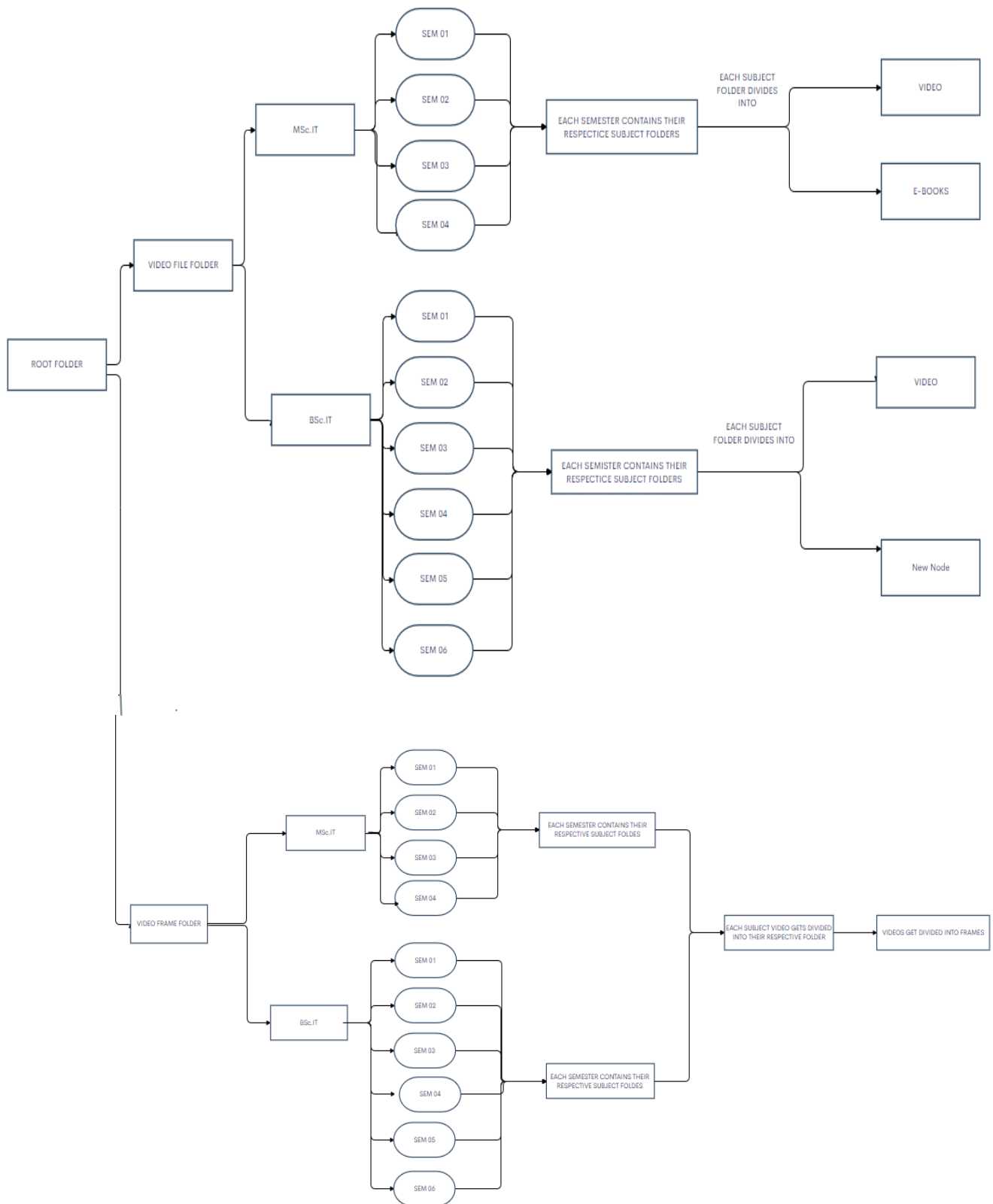


FIGURE 1 DATABASE MODEL

### 3.5. System diagram

#### 3.5.1. Basic Module

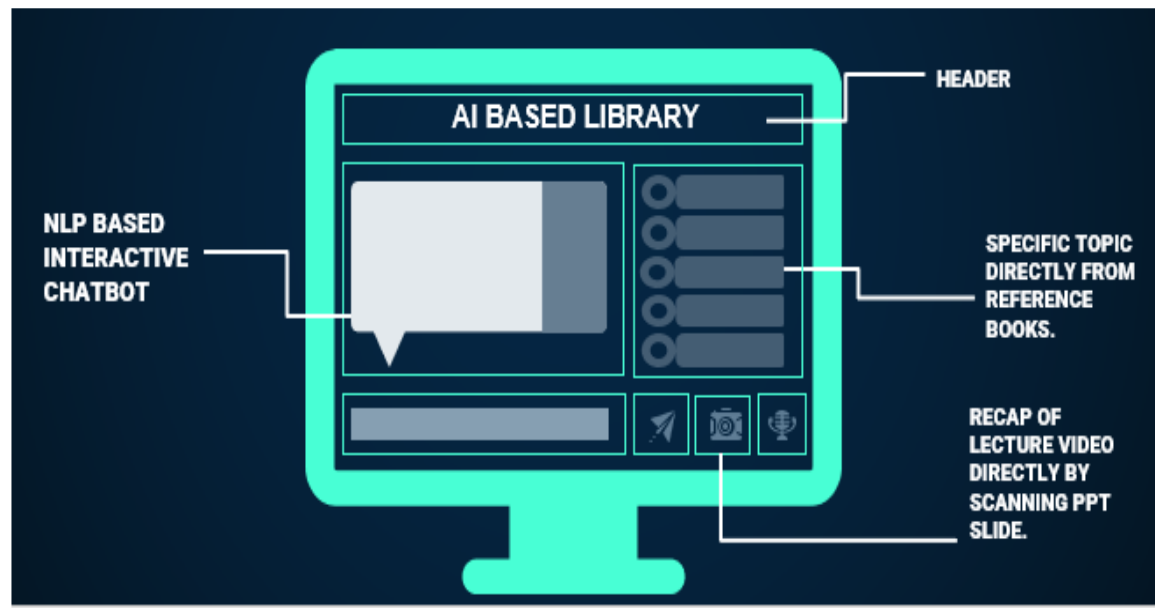


FIGURE 2 : BASIC MODULES

#### 3.5.2. Use Case Diagram

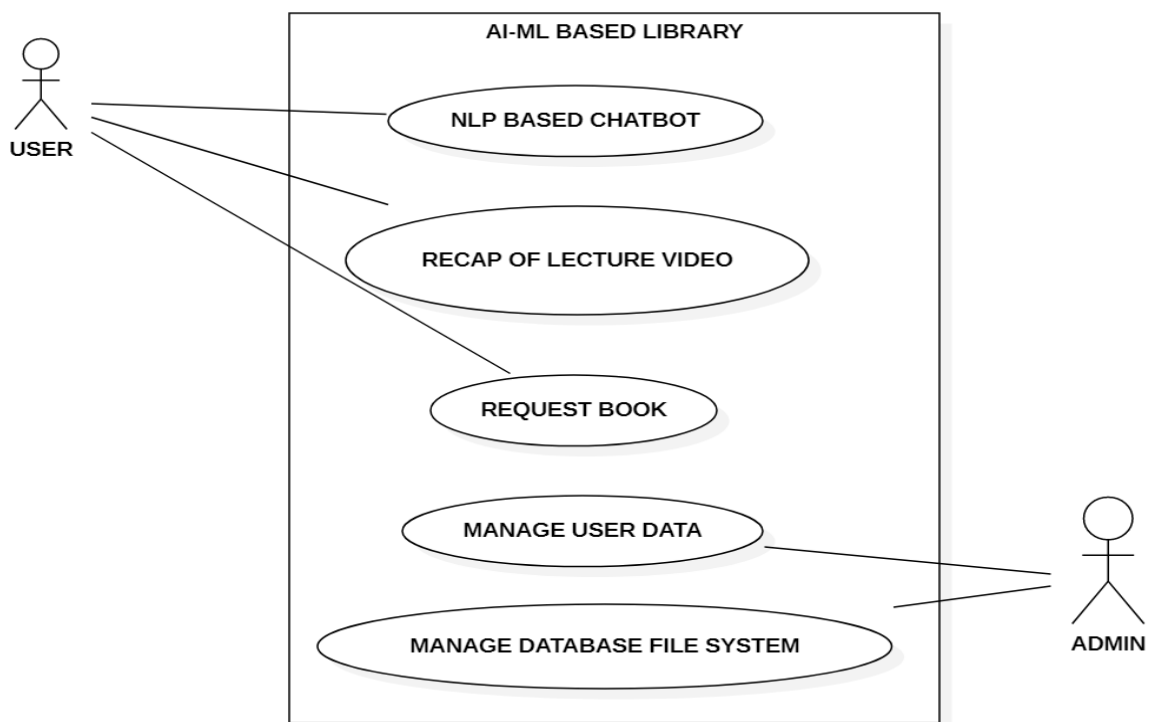


FIGURE 3 USE CASE DIAGRAM

## **CHAPTER 4: METHODOLOGY**



## 4.1. Introduction

Methodology used for this project comprised of various combination of general as well as ML based specific techniques to get the features of this project working. For Video search feature OpenCV based pre-trained models have been used which are able to search the video based on the referenced frame, whereas for the nlp based chatbot for assisting purpose is a mix of traditional programming techniques such as regex and classification-based Information Extraction methods.

## 4.2. Block diagram/Flowchart of proposed methodology

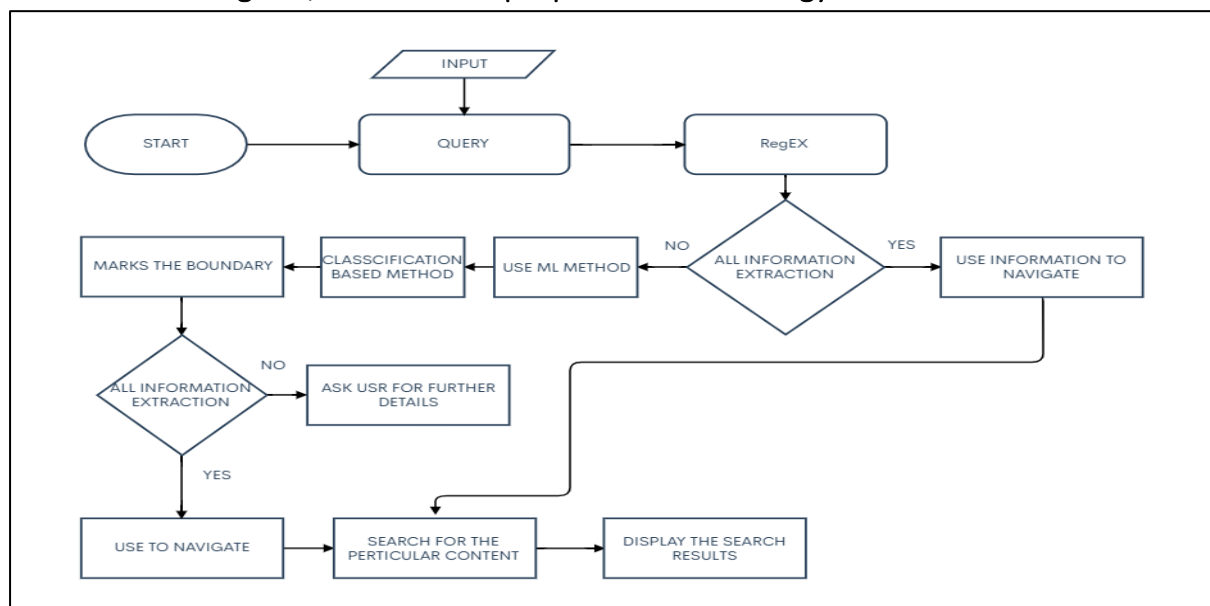


FIGURE 4 :FLOWCHART FOR NLP BASED CHATBOT

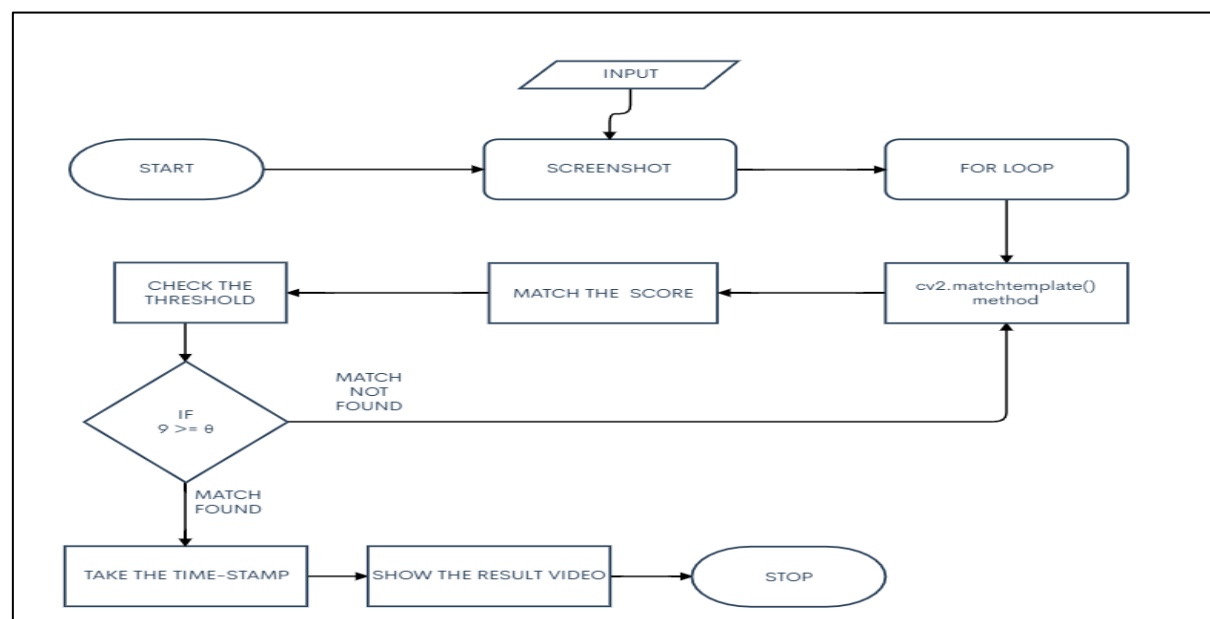


FIGURE 5 FLOWCHART FOR VIDEO SEARCH

### 4.3. Description of methodology used

#### **NLP Based Chatbot**

For NLP based chatbot the methodology we have used to extract information from a given query is Classification Based Extraction Methods:

The basic idea is to present information extraction problem as that of the classification. Boundary detection using classification model, this system consists of two distinct phases:

- Learning

In the learning phase our system uses a set of labelled documents to generate models Which we can use for future predictions.

- Extracting

The extraction phase takes the learned models and applies them to new unlabeled Documents using the learned models to generate extractions. This model aims to generate the boundaries (start boundary and end boundary) for the special information.

The method formalizes the IE problem as a classification problem. It is aimed at detecting the boundaries (start boundary and end boundary) of a special type of information. For IE from text, the basic unit that we are dealing with can be tokens or text-lines in the text. Then we try to learn two classifiers that are respectively used to identify the boundaries. The instances are all tokens in the document. All tokens that begin with a start-label are positive instances for the start classifier, while all the other tokens become negative instances for this classifier. Similarly, the positive instances for the end classifier are the last tokens of each end-label, and the other tokens are negative instances.

There are two classifiers – one to identify starts of target text fragments and the other to identify ends of text fragments. Here, the classifiers are based on token only (however other patterns, e.g., Syntax, can also be incorporated into). Each token is classified as being a start or non-start and an end or non-end. When we classify a token as a start, and also classify one of the closely following token as an end, we view the tokens between these two tokens as a target instance.

In the extracting stage, we apply the two classifiers to each token to identify whether the token is a “start”, “end”, neither, or both. After the extracting stage, we need to combine the starts and the ends predicted by the two classifiers. We need to decide which of the starts (if there exist more than one starts) to match with which of the ends (if there exist more than one ends).

To summarize, this IE classification approach simply learns to detect the start and the end of text fragments to be extracted. It treats IE as a standard classification task, augmented with a simple mechanism to combine the predicted start and end tags. Experiments indicate that this approach generally has high precision but low recall. This approach can be viewed as that of one-level boundary classification.

### **Video Search Feature:**

To implement video search feature using screenshot various methods have been implemented in various different ways with little bit here and there in many research papers using Machine Learning method as well as using traditional methods as well. The Methodology used in this project for this feature is more of a using traditional way were using various image processing methods and mathematical distance as well as other metric proves to be much faster, accurate and efficient method to do so. As this project focuses on videos of lectures which generally includes ppts and some writing work, if used for machine learning the methodologies suffer to find much attribute or features to learn to distinguish between different frames due to the lack of much difference in subsequent frames of the video.

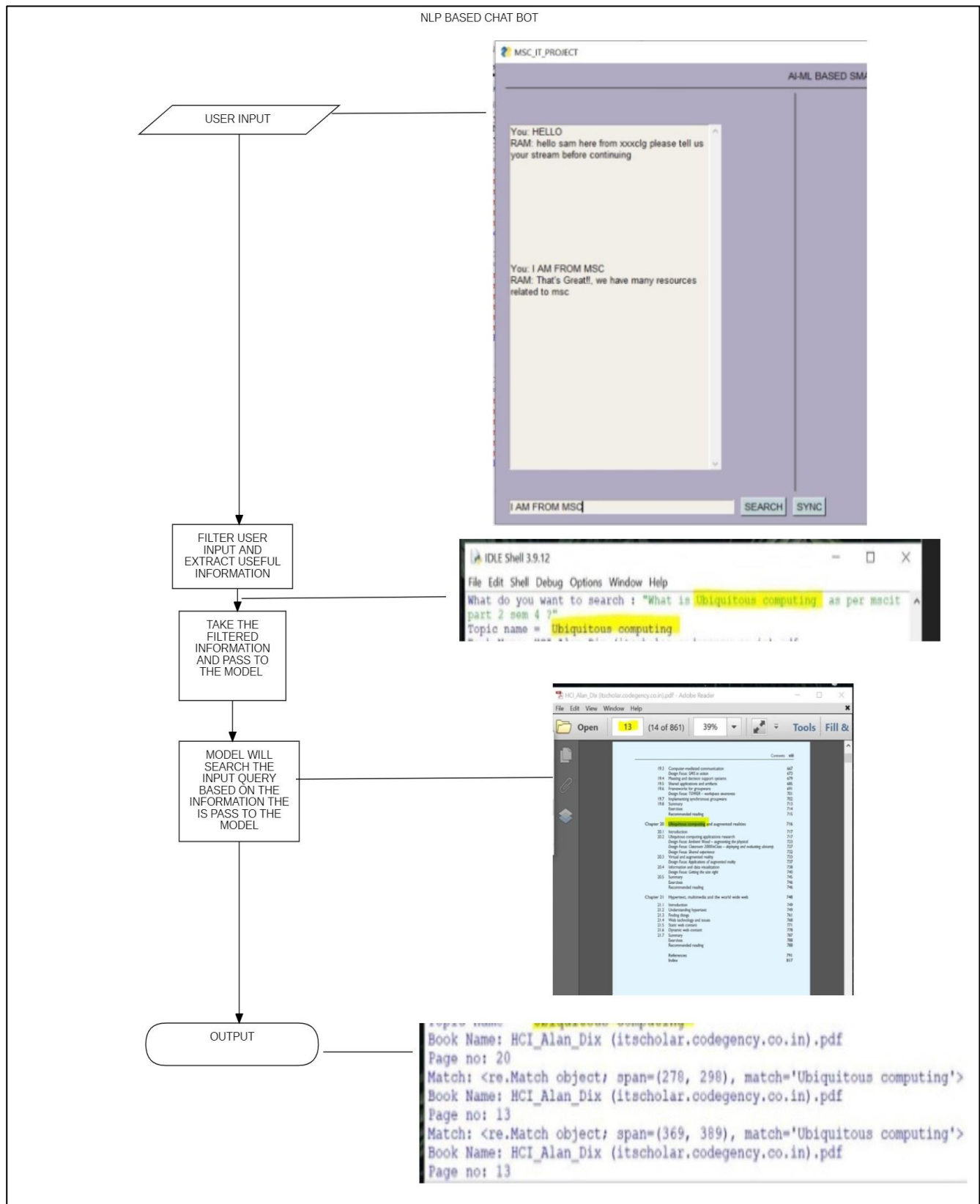
So, the solution that we came on to was to use more of a traditional method that was able to look for a match rather than to look for a difference, by changing the perspective of the methodology we gained much better results in post testing of the methodology in comparison with the older one. To implement this search methodology using feature match technique we went with OpenCV's one of the most famous method know as match template. This feature has been very popular in many computers vision projects as well as pre-processing method of various machine learning techniques.

Matching template takes to template and returns a result which a match score telling us that this much the given two templates match. One of the most interesting features of this method is that identifying text in templates is very fast and accurate which helped in our project as well as most of the ppts generally has a lot of text in them which match template could easily look for and gave result We set our threshold of minimum match as high as possible so that we can have as accurate results as possible.

So, the whole process of video searching feature goes on first creating different frames out of a video and then using a loop to iterate all the frames with the given screenshot to check which one matches and a small code to highlight which are the match has found. Once the highest score is found out then we use the name of the frame to present video from that second to the user, as the name of every frame includes the second from which the frame has been taken.

## **CHAPTER 05: RESULTS, ANALYSIS AND DISCUSSION**

### 5.1. NLP Based Chatbot:



**FIGURE 6: NLP BASED CHATBOT OUTPUT**

## 5.2. Video Search Feature:

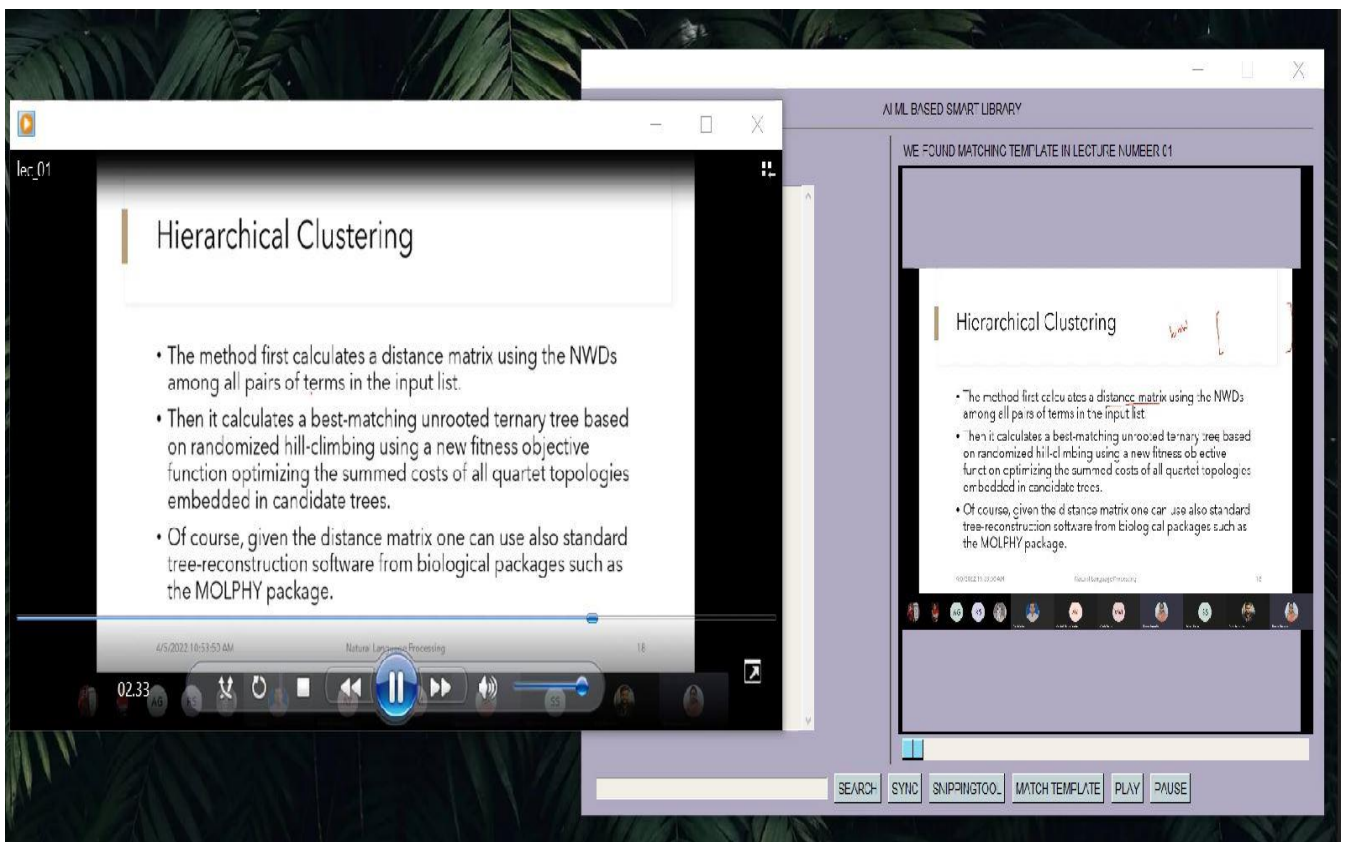
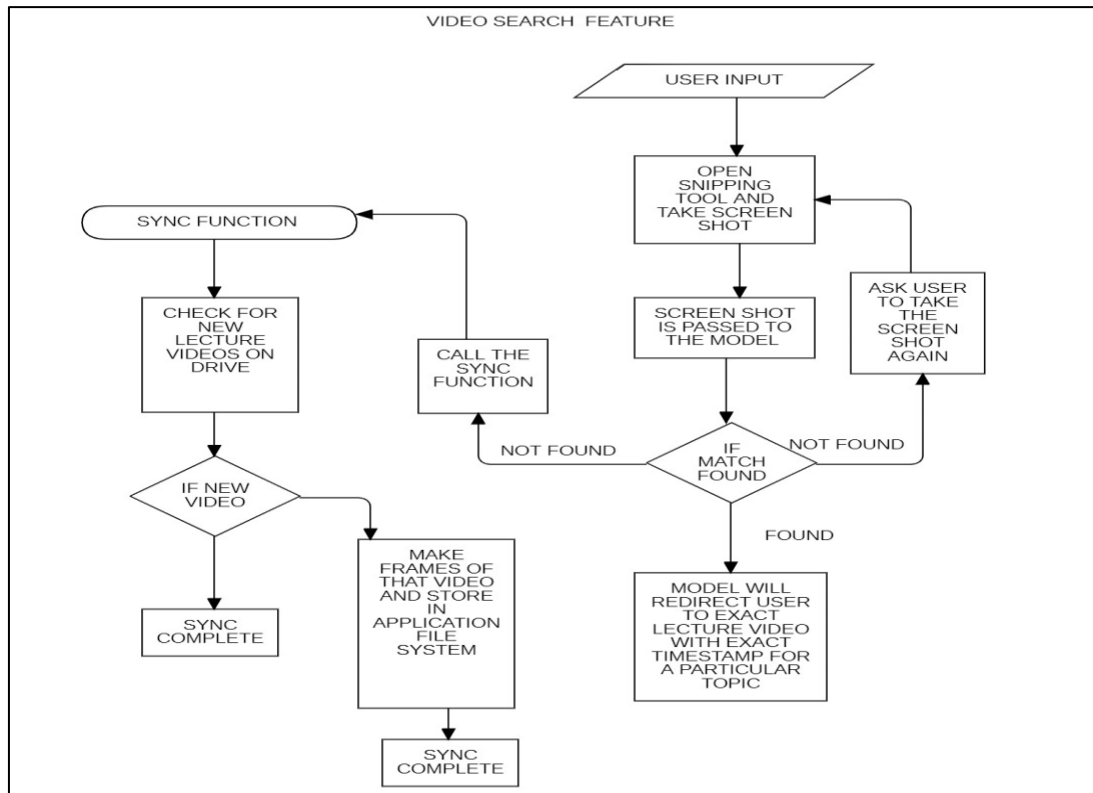
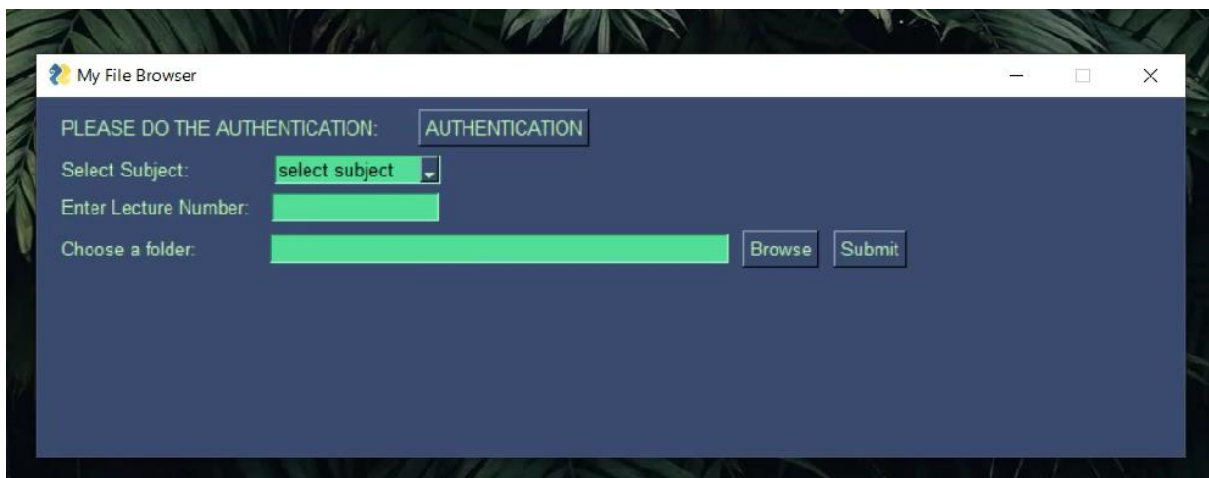


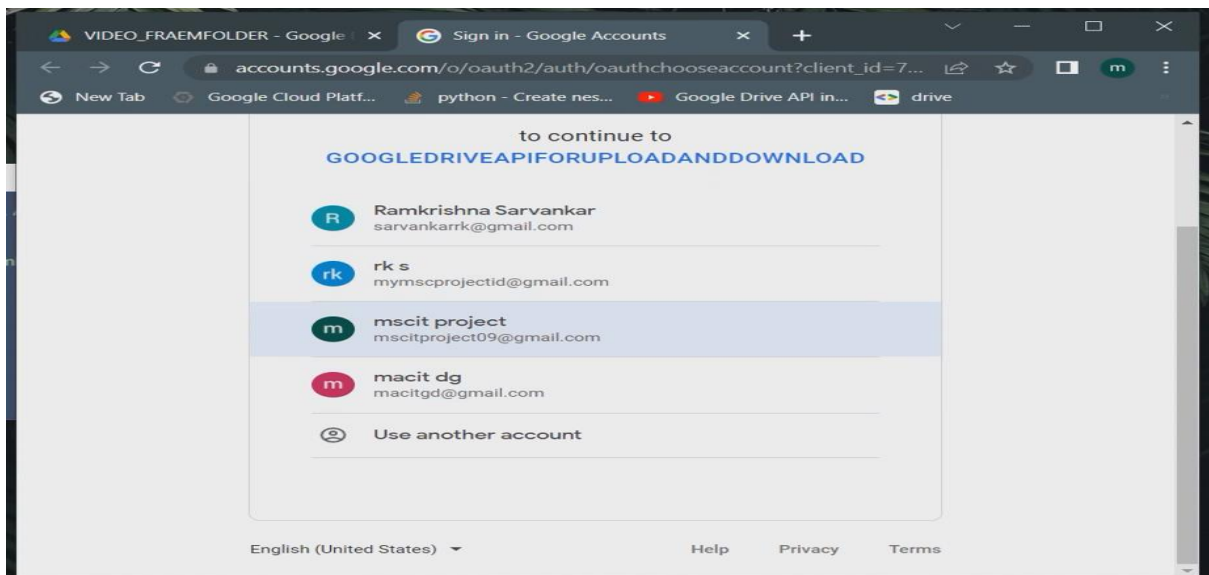
FIGURE 7: VIDEO SEARCH MODULE OUTPUT

### 5.3. Admin Side Module:

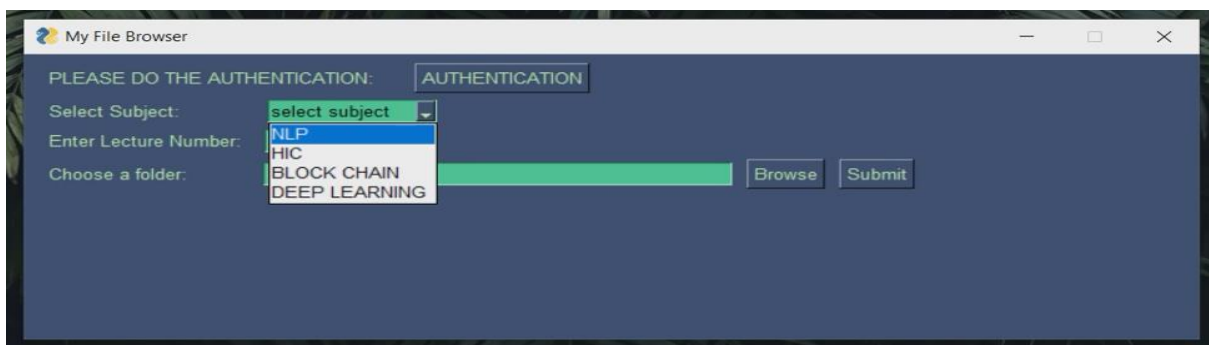
#### STEP 01 – OPEN ADMIN SIDE MODULE



#### STEP 02 – COMPLETE THE AUTHENTICATION USING YOUR GMAIL ID



#### STEP 03 – THEN SELECT THE SUBJECT



STEP 04 – ENTER LECTURE NUMBER AND SELECT THE LECTURE VIDEO FILE USING BROWSE BUTTON AND UPLOAD THE FILE ON DRIVE USING SUBMIT BUTTON



The screenshot shows a web application titled "My File Browser". It has a dark blue background with white text. At the top, it says "PLEASE DO THE AUTHENTICATION:" followed by a button labeled "AUTHENTICATION". Below this, there are three input fields: "Select Subject:" with a dropdown menu showing "NLP", "Enter Lecture Number:" with a text input showing "15", and "Choose a folder:" with a text input showing "P:/msc\_project/dataset/msc\_it/sem\_04/nlp/video\_frame". To the right of the folder input are two buttons: "Browse" and "Submit".

STEP 05 – THEN YOU CAN SEE THAT LECTURE VIDEO IS UPLOADED ON TO THE DRIVE

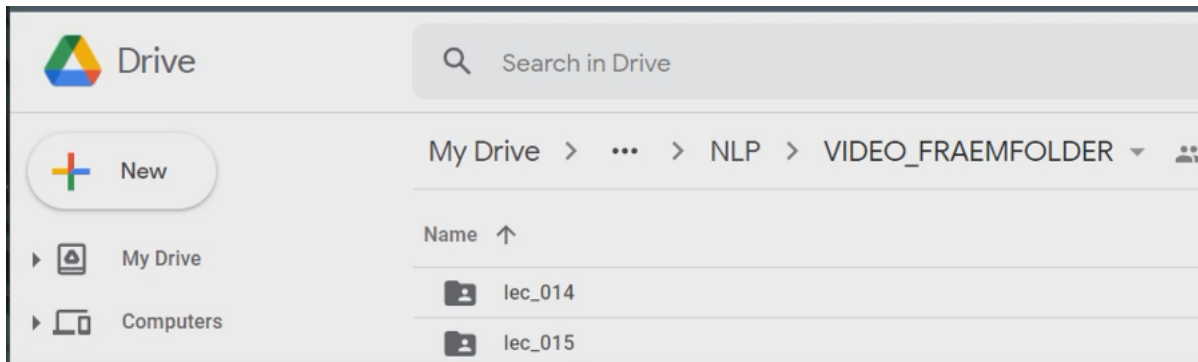


FIGURE 8:ADMIN SIDE MODULE OUTPUT



Sr. No	Event	Associated Activity	Expected Output	Outcome
1	User Authentication	Valid Login Input	Should Succeed in Login	Successful
2	Admin Authentication	Valid Login Input	Should Succeed in Login	Successful
3	User Authentication	Invalid Login Input	Should Fail to Login	Failed to Login
4	Admin Authentication	Invalid Login Input	Should Fail to Login	Failed to Login
5	Chat Window	Proper Input is Given	Should give a Reply by using NLP based Model	Successful Reply
6	Chat Window	Improper Input is Given	Should Reply with "I did not understand"	Successful Reply
7	Chat Window	Repetitive input	Should same replay all the time	Successful Reply
8	Video Search module	Proper Input is Given	Should redirect user to the lecture video with exact timestamp	redirect user to the lecture video with exact timestamp
9	Video Search module	Improper Input is Given	Ask user to give the input again	Successful Reply
10	Video Search module	Improper Input is Given	Sync application with the drive data	Sync complete

**TABLE 1:Event Table**

## **CHAPTER 06: CONCLUSION, LIMITATIONS, AND FUTURE WORK**

## 5.1. Conclusion

With all the research and development that went in designing and developing the AI/ML based Library has concluded into various possibilities and understanding which has the potency to make the learning path of the user much for easy to travel and intuitive to be a part of.

The understandings that this project concluded in are as follows,

- 1) Students are facing a lot more subtle problems than the ones that are in highlight, which is affecting the student's learning capabilities in a drastic manner, and it requires the most amount of attention and work.
- 2) Problems are there in various domains and sectors of a student's learning path but most of the problems that are there, lies in the ways and manners in which the student is approaching the various methods of learning.
- 3) A constant companion is required with most amount of knowledge and resources to help the students in dealing with their learning problems.
- 4) Artificial Intelligence can fulfill some of void of a companion if implemented properly.
- 5) Time of a student should be invested in learning and not finding the resources to learn.
- 6) Internet is the best place for learning in terms of resources and teaching if served precisely.
- 7) Artificial Intelligence can help in filtering the internet for a student as per student.

The possibilities that this project concluded in are as follows,

- 1) Students can now have access to various resources especially as per their curriculum with efforts of few clicks.
- 2) Video referencing for learning especially topic wise is now a single click away.
- 3) Students can have now much more time to invest in topics that require much more attention without worrying for deadlines.
- 4) Finding answers from multiple reference books and even directly from lecture videos is now a possibility without going to search for each and every book and video manually.
- 5) A single platform is now available for students to look for each and every kind of resources for any topic.

Teachers and students with the help of AI/ML based library can reach new heights of learning and can shape the path of learning in much better way, which will cater the students' needs and giving them ease with lots of time to spare it elsewhere.

## **5.2. Limitation**

Limitation of this project totally depends on the limitations that AI/ML has in terms of computing and inferences that it can provide and automation that it can do with the help of newer technologies and development in this area.

Especially talking about the limitations of the project, there are various limitations that it faces in terms of performance, reliability, accuracy and all other such factors that an AI/ML based project has to face.

Limitation of the AI/ML based Library are as follows,

- 1) As the data increases on the server, the project might face issues with the speed at which it provides the results but there are various methods that can help in solving the performance related issues.
- 2) Reliability of any AI/ML based project is always in question and nothing is different in this project as well, were this project as the reality dictates that because of its predictive nature it can sometimes depending on situation, can lead to scenarios where the results might not be as accurate as it should be, but the methods that are implemented are very less on probability side and more on static computation side of the AI/ML domain, which promises good reliability in terms of its results and accuracy.
- 3) Due to the project being heavy on resources which can be computation as well disk consumption, it will not be available in offline mode which sometimes and in areas with less internet capability can become an issue or roadblock in implementation of the project.
- 4) If the resources on which the whole project is running such as pdf books and recorded lectures are not provided and well managed, this project will not be able to help the student the way it promises to its users.

## **5.3. Future Work**

Future work on the project has a huge list which includes implementation of various new features for accessing a greater number of resources and in various other ways.

The future work on this project are as follows,

- 1) Integrating YouTube videos with the recorded lecture videos for much more diversity.
- 2) Presenting google search results filtered in different ways for the topic searched.
- 3) Building a collaboration module for doubts and discussions on searched topic and for help if needed.
- 4) Building APIs for the library to access the features of library directly from other services and applications.
- 5) A constant effort to increase the efficiency at which student tries to use to the application

# Summary

In this project documentation we saw the scenario around the learning of the students followed by the problems a student faces while using the resources and then we saw how the research and development of the AI/ML library will help the students to solve their problems.

So, in introduction part we saw that students basically learn from either the lectures or the reference books or materials, especially in the scenarios of curriculum-based studies.

Then we moved on to the various problems that students are facing to cover their revisions during exam times and during usual study times as well which included, the problems related to navigating the resources and then accessing and then while using them, all of them challenged the students in some or the other ways, which was getting reflected from students increased time consumption in searching the resources rather than studying them.

After that we introduced our research and design of AI/ML based Library and how its implementation of various new technologies will try to help the students in need of them.

We presented you with various technologies and the ways of implementing them, the one we choose to implement and stated all the reasons for doing so which included various methodologies for implementing the chatbot and video search module as well.

We sketched some of the important diagrams of the various phases of development and interaction models to understand the working and flow of execution in the project.

Moving on various screenshots of the ui of the application is been presented with its uses as well.

Then we concluded our project with various learning and possibilities we found in the research as well development journey, followed by limitations that our project comes with.

## References

- I. **Ciravegna F.** (LP)<sup>2</sup>, an adaptive algorithm for information extraction from Web Related Text [Conference] // Workshop on Adaptive Text Extraction and Mining held in conjunction with 17th International Joint Conference on Artificial Intelligence . - Seattle, USA. : [s.n.], 2001.
- II. **Finn A.** A multi-level boundary classification approach to information extraction. [Report]. - Dublin : University College, 2006.
- III. **Finn A., & Kushmerick, N.** Information extraction by convergent boundary classification. [Report]. - San Jose, USA : AAAI-04 Workshop on Adaptive Text Extraction and Mining., 2004.
- IV. **Jie Tang Mingcai Hong, Duo Zhang, Bangyong Liang, and Juanzi Li** Information Extraction: Methodologies Applications [Journal]. - Beijing : Tsinghua University, 2008.
- V. **Kristjansson T. T., Culotta, A., Viola, P. A., & McCallum, A.** Interactive information extraction with constrained conditional random fields. [Conference] // In Proceedings of AAAI'04. - 2004. - Vols. pp 412-418.
- VI. **Li J., & Yu, Y.** Learning to generate semantic annotation for domain specific sentences. [Conference] // In Proceedings of the Knowledge Markup and Semantic Annotation Wrokshop.. - Vlctoria, BC : [s.n.], 2001.
- VII. **Li X., & Liu, B** Learning to classify texts using positive and unlabeled data. [Conference] // In Proceedings of International Joint Conference on Artificial Intelligence.. - 2003. - Vols. pp 587-592.
- VIII. **Muslea I.** Extraction patterns for information extraction tasks: A survey. [Conference] // In Proceedings of AAAI-99: Workshop on Machine Learning for Information Extraction. . - Orlando : [s.n.], 1999.
- IX. **Riloff E.** Automatically Constructing a Dictionary for Information Extraction [Conference] // Eleventh National Conference on Artificial Intelligence. . - 1993. - Vols. pp 811-816.
- X. **Soderland S., Fisher, D., Aseltine, J., & Lehnert, W** CRYSTAL: Inducing a Conceptual Dictionary [Conference] // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. - 1995. - Vols. pp.1314-1319..
- XI. **Tang J., Li, J., Lu, H., Liang, B., & Wang, K.** iASA: Learning to Annotate the Semantic Web [Journal]. - New York, USA : Springer Press. , 2005. - pp. 110-145 : Vol. Journal on Data Semantic IV.

