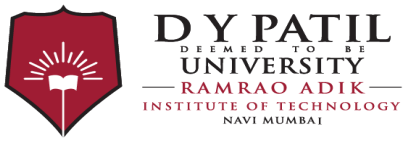


Ramrao Adik Institute Of Technology
(D Y Patil deemed to be University)
Dr.D.Y.PatilVidyanagar,Sector-7,Nerul,Navi Mumbai400706.

External Feedback Analysis

Internship Project Report

Submitted in partial fulfillment of the requirements of

Software Development Group (SDG)-RAIT

by

Kunal Kamble [19CE5503]
Atharva Shanware [18CE7013]
Akshay Shringarpure [18CE8002]
Ajinkya Surve [18CE8015]
Rohan Hasabe [19CE1071]

Supervisor

Dr. Leena Ragha
Mrs. Saguna Ingle

AUGUST-2021

RamraoAdik Institute of Technology

(D Y Patil Deemed To be University**)**

Dr.D.Y.PatilVidyanagar,Sector7,Nerul,NaviMumbai400**706.**

CERTIFICATE

This is to certify that,the internship project titled

"External Feedback Analysis"

is a bonafide work done by

Kunal Kamble [19CE5503]
Atharva Shanware [18CE7013]
Akshay Shringarpure [18CE8002]
Ajinkya Surve [18CE8015]
Rohan Hasabe [19CE1071]

and is submitted in the partial fulfillment of the requirement for the Internship Programmer under SDG-RAIT

Co-supervisor	Supervisor
(Name of Co-supervisor)	(Name of Supervisor)

Internship Co-ordinator	Internship Co-ordinator	Vice-Chairman SDG-RAIT	Chairman SDG-RAIT
(Mrs.Smita Bhoir)	(Mrs.Dhanashri Bhosale)	(Prof.LeenaRagha)	(Prof.M.D.Patil)

Title : **External Feedback Analysis**

Class of work : **Software**

Authors : Kunal Kamble [19CE5503]

Atharva Shanware [18CE7013]
Akshay Shringarpure [18CE8002]
Ajinkya Surve [18CE8015]
Rohan Hasabe [19CE1071]

Organization : **Ramrao Adik Institute of Technology, Navi Mumbai**

INDEX

List of Figures

Abstract

External feedback Analysis system is used to get the feedback from the alumni,external examiner,industry expert and parents.It generates the reports for different domains based on this feedback.The Staff and the Students will be provided with separate usernames and passwords in order to check the results.We will analysis the feedback data using natural processing language(NLP). NLP refers to the branch of artificial intelligence(AI) which is concerned with giving computers the ability to understand text and spoken words.The results will be displayed in the form of charts and graphs.

Alumni feedback will help students to know about the challenges faced in the industry, skills developed in RAIT, areas where RAIT can help students for placements.External feedback will help students know about the overall experience, basic understanding of course, communications skills**.** Industry feedback will help students know the latest industrial trends, courses and skills required.PTM feedback will help the staff as well as parents to understand their ward's performance, study pattern.

Chapter 01

Introduction

1.1. Motivation

Feedback is an important component of effective learning and an important pillar for development of any institution.It can improve student's confidence ,self-awareness and enthusiasm in learning. Although, all the data regarding the various domains is available, it is not been presented and saved in a way that can help the students,faculty and administration to improve along the changes in the industry.This is where our project plays an important role.We collect the data regularly and analyze it using NLP models and present the output in the form of pie-charts, graphs and histogram that are much better to observe.

1.2. Purpose and Problem Definition

The project is devised taking into account the present world needs as well as the current issues faced in gathering information from different domains.It enables college administration, faculty, students, parents to manage and analyse the data obtained through feedback forms.This would help students in understanding the current trends of the industry while staying at home and without the need of any hard copies of forms and documents.It would help parents to monitor their ward's performance.It would help faculty understand the latest topics required to be taught. Overall it would be very helpful for the entire organization to excel in their particular domains by analyzing the feedback which will be displayed in the form of pie-charts,graphs and histograms.

1.3. Scope

Based on various research, the proposed External Feedback Analysis project would be able to provide an analysis on details filed by the Alumni, Industry Men, External Examiner and Parents through the forms.The proposed system is designed to reduce the time and save the efforts of the faculties from maintaining huge amounts of records. At the time of the feedback generation it applies the proposed algorithms and methodologies for generating feedback of a specific category. After this a whole collective report is generated with pie-charts and graphs showing the collective rating of a particular field and a collective opinion is provided on the basis of analysis done on the

basis of details provided in the form about any particular field such as course structure, subject's topics etc.

Chapter 02

2.1. Objectives

Feedback is an crucial aspect of any given organisation and its growth.Organizations collect feedbacks from relevant public so that they can make improvements to their current methods of operation and try to solve as many issues as possible to engage their clients.But feedback forms are mostly in text formats and thus it will be a tedious procedure to process all this data by a person or a team of people and extract some meaningful information.To simplify this process,we use the concept of machine learning so that a customized program can process the data submitted in the feedback forms and generate statistical data that can be visually represented in the form of charts and graphs to deliver concise and relevant information to the organisation that they can easily act upon.

Although this project includes an array of software domains such as natural machine learning, web development and database management,the primary objective is to develop a keyword identification software using natural language processing.Using the textual data submitted by professionals,industry experts and other visitors to the feedback form pages,we first extract the keywords from the data and compare it to the listed keywords in the models.

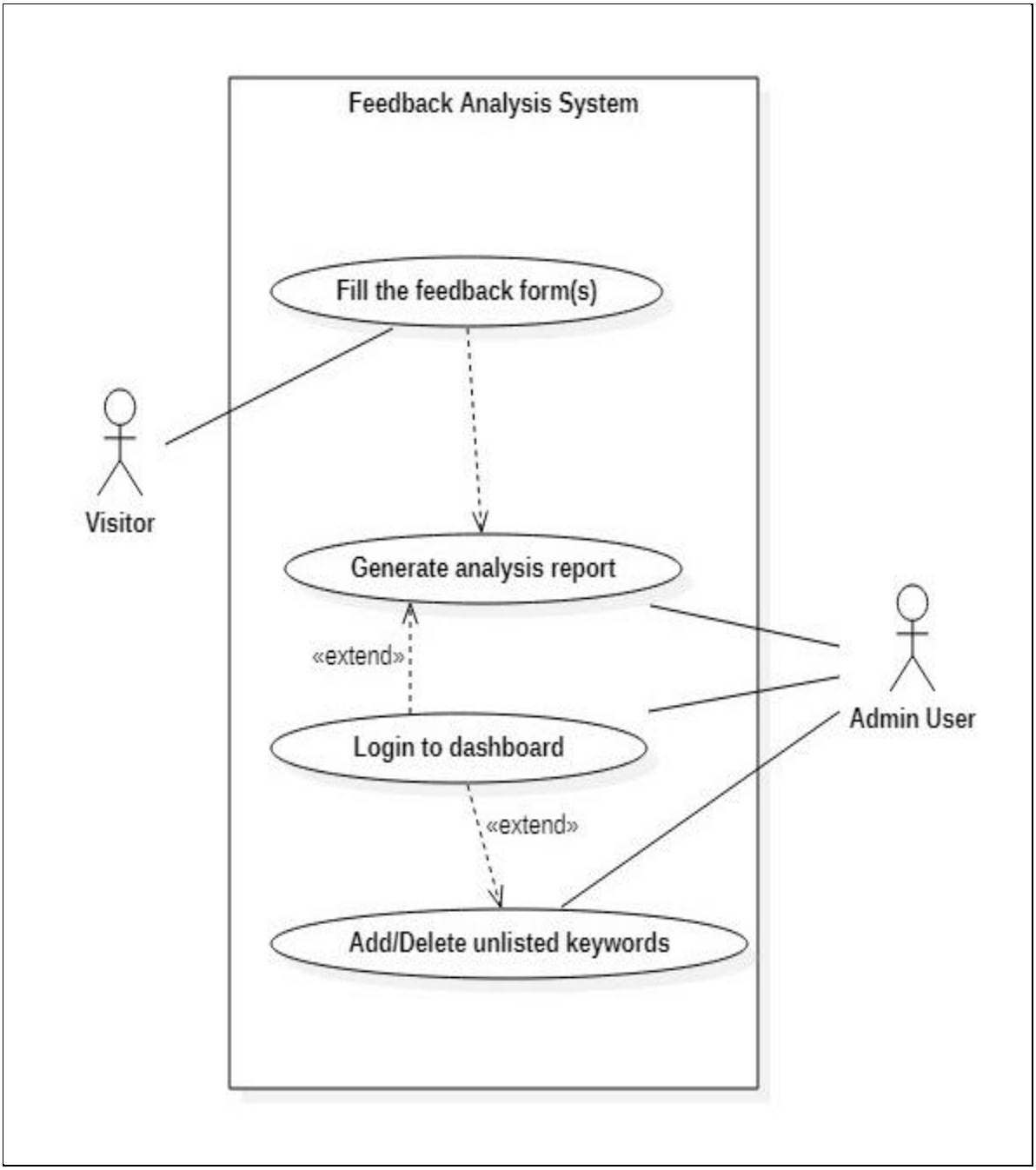
Based on the keywords detected, the software would display a histogram based on the number of occurrences of those keywords in the responses and recognise trending keywords among visiting users.

The analysis report page would also display pie charts based on other questions in feedback form that are multiple choice-based in nature.

Due to the possibility of detected keywords not existing in the models,we wish to employ a dynamic approach where the administrative user would have the option to add or remove unlisted keywords to the models.

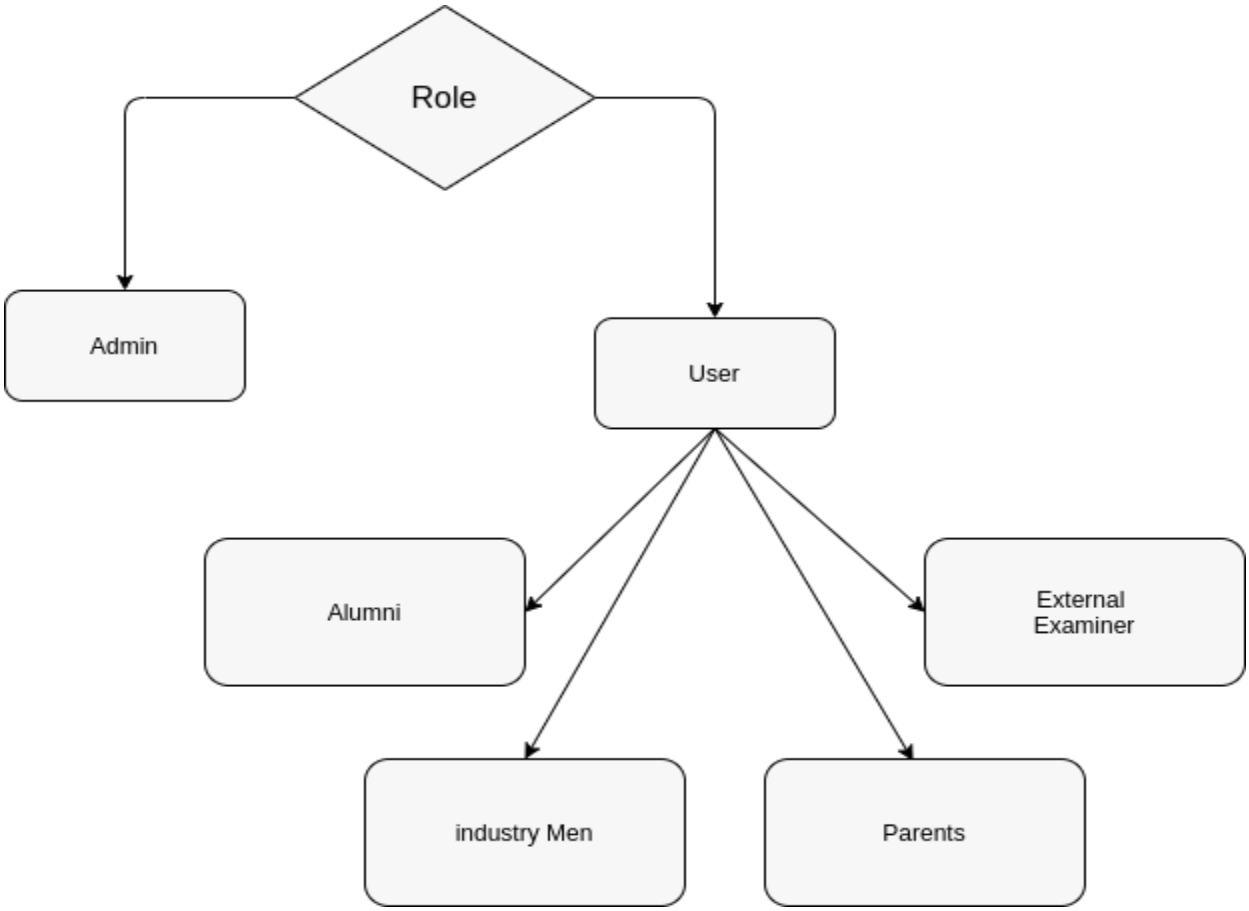
2.2. Structure of the proposed system

- **Use Case Diagram**

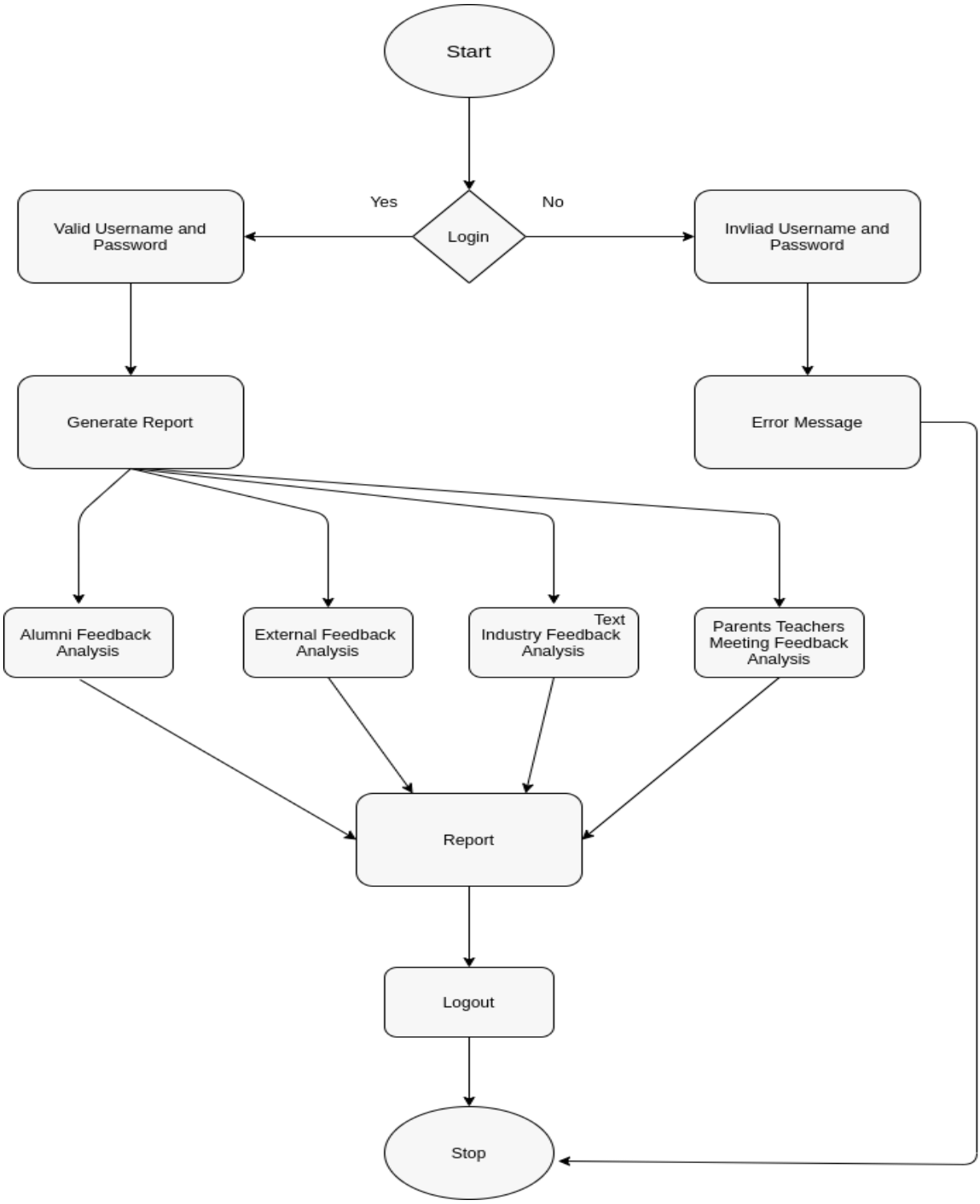


fig(2.2.1) : Use Case Diagram of the proposed system

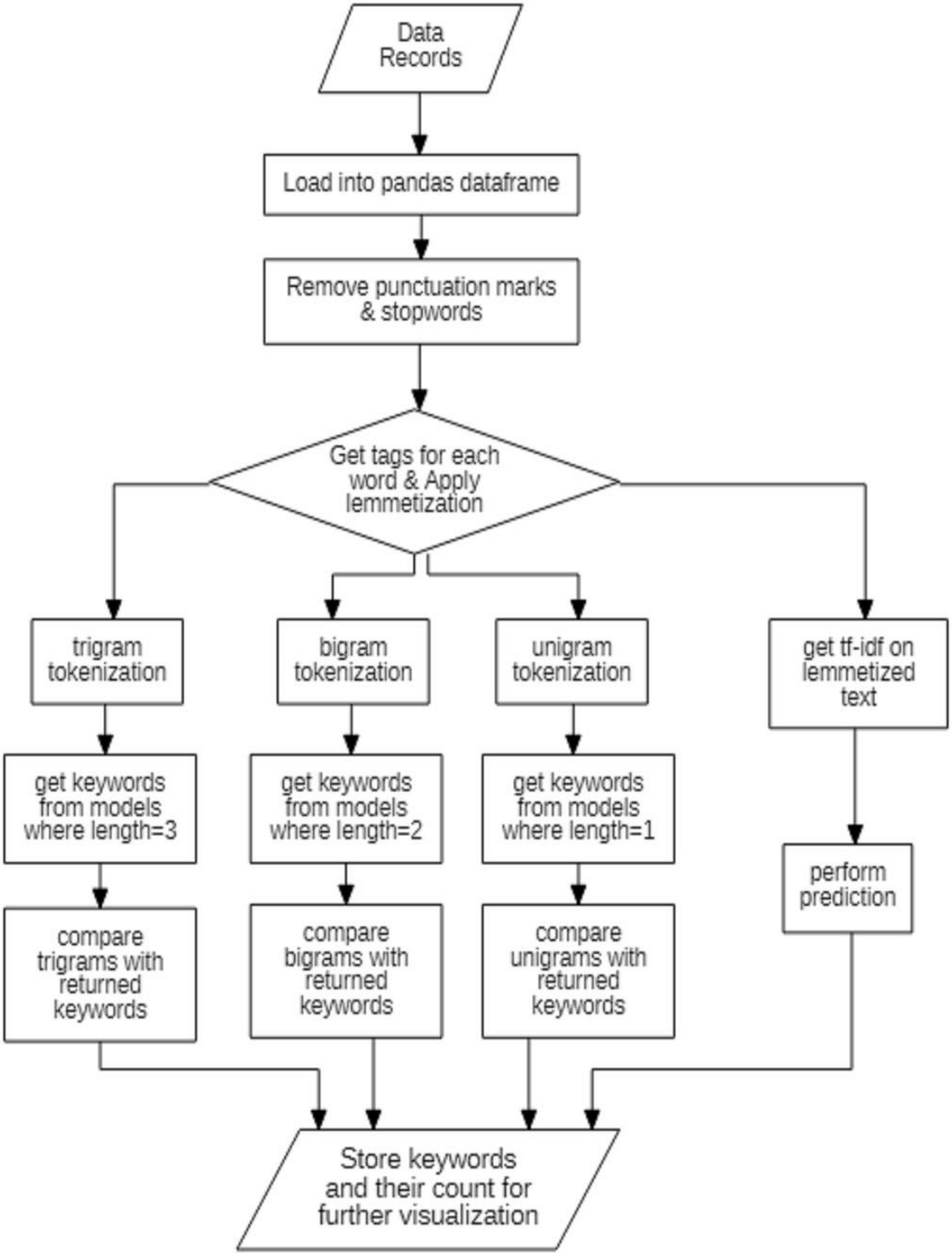
- **Activity Diagram**: An Activity diagram is a behavioral diagram i.e. it depicts the behavior of a system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.



fig(2.2.2): Activity Diagram for Roles



fig(2.2.3): Activity Diagram for Admin



fig(2.2.4):Flowchart of Keyword Identification algorithm

2.3. Functioning :

- Visiting users, i.e. individuals visiting the website to fill the feedback form only have access to the different kinds of feedback forms such as external feedback, alumni feedback, industry feedback and PTM feedback. They can visit the homepage of the website, select the appropriate form button and upon clicking, they will be redirected to the form. Once all the required fields are filled and the form is submitted the page refreshes and visitors are allowed to submit another response.
- Authenticated users(admin) who have the login credentials can sign in through the login page. Upon successful login,they are directed to the dashboard/profile page. Admin can simply select the required feedback category from a drop-down menu and click on “Generate Report” to view the analysis report generated based on the existing data records for the specified category.
- On the report page, you can find the button ''Unlisted Keywords'' at the navigation bar in the top right corner. Upon clicking this button, admin will be directed to the list of detected keywords not present in the preconfigured list used by the backend program. Then the admin can decide whether the keywords need to be added to the models or if those keywords are unnecessary and can be deleted.
- Once the admin is done viewing the reports,he/she can return to the dashboard/profile page and log out. After successfully logging out,the admin will be directed to the home page.

2.4. Novelty of Approach :

The ML model that we have designed employs natural language processing to extract keywords from the textual input provided in the form fields by visiting users.The algorithm first cleans the text by removing punctuation marks and stopwords, then segregates the detected keywords based on the number of words present and compare it with the list of keywords predefined in the program.

We have also added the functionality to detect unlisted keywords and allow admin users to personally decide if those keywords are legitimate and need to be added to the pre-existing list of keywords or can just be ignored and removed.

These deleted keywords would not be permanently blacklisted as there is a possibility that a certain keyword is not trending today but would be relevant in the coming future.

For visual representation of the processed data, we opted for dynamic charts and graphs instead of using static images of the data visualization can be done at the backend. This approach benefits this project as it provides better understanding of the statistical data procured from the data records and will also give the webpage a more high-end and professional look.It was observed that the generation of report will be delayed in time as the number of records increase, but this issue can easily be overcome by using high-end servers with sufficient computational power to generate visualization objects instantaneously.

Chapter 03

3.1. Research methodologies and technologies:

Natural Language Processing (NLP) is a process of manipulating or understanding the text or speech by any software or machine. An analogy is that humans interact and understand each other's views and respond with the appropriate answer. In NLP, this interaction, understanding, and response are made by a computer instead of a human.

NLP libraries in python

- **Natural language Toolkit (NLTK)**: NLTK (Natural Language Toolkit) is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.
- **TextBlob**: *TextBlob* is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- **Scikit-learn**: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Dataset for training the ML model:

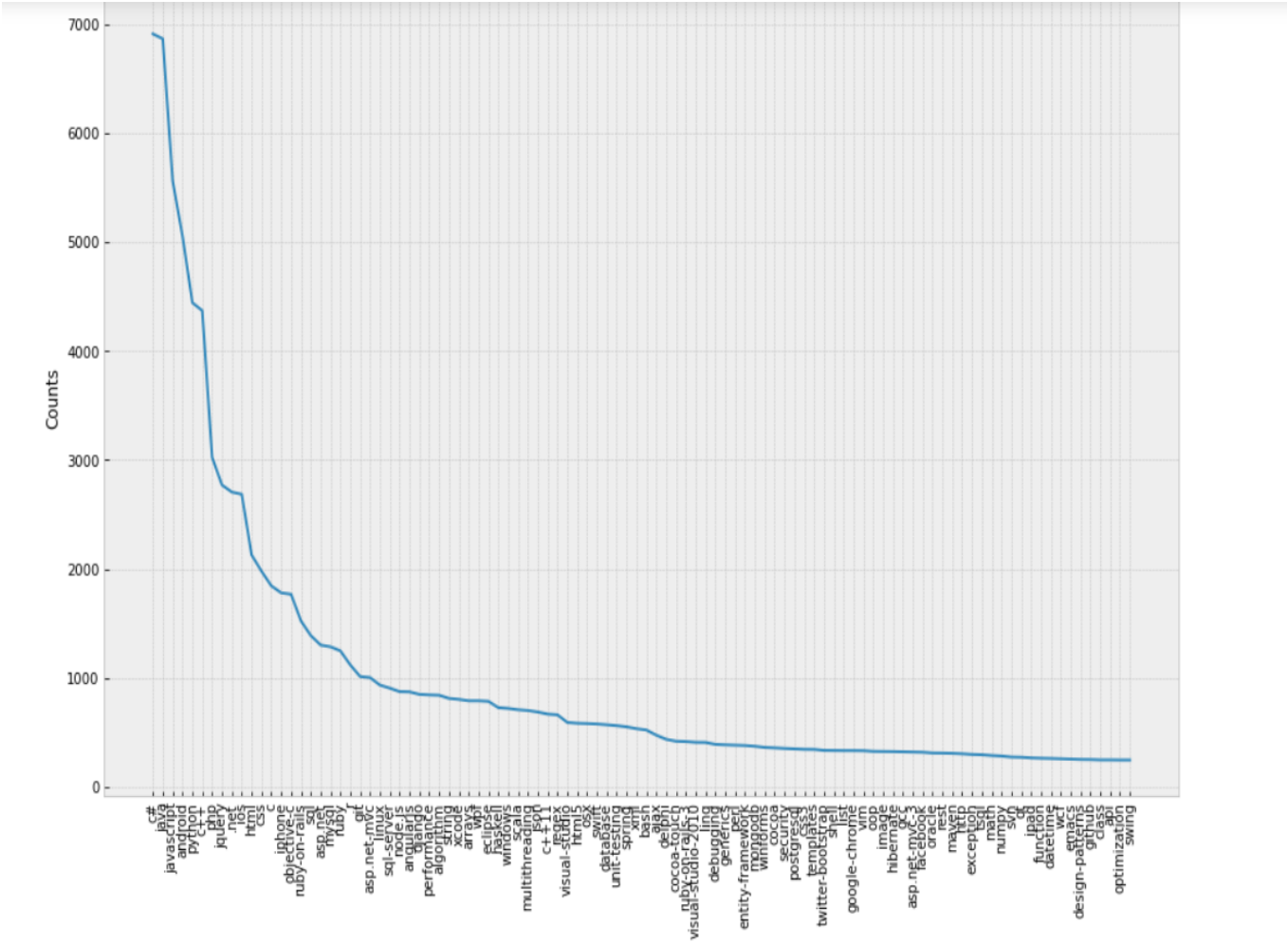
Dataset with the text of 10% of questions and answers from the Stack Overflow programming Q&A website.
This is organized as three tables:

- Questions contain the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.
- Answers contain the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
- Tags contains the tags on each of these questions

Preparing the data:

Before training the NLP model, the dataset needs to be pre-processed to highlight the important features of the data. The steps for data cleaning are:

- **Data cleaning**: The initial dataset consists of 1264216 rows of data consisting of questions and tags associated with each question. In data cleaning all the punctuation marks present in the text data are removed along with the stopwords. The stopwords are a list of words that are very common but don't provide useful information for most text analysis procedures. To keep the consistency within the tags associated with the text, only the questions with score more than 5 points are considered, then top 100 tags are chosen according to their frequency. This reduces the row count to 63167.



fig(3.1.1): Top 100 tags and their frequencies

- **Tokenization**: Tokenization is a process of parsing the text string into different sections (tokens). It is necessary since the computer programs understand the tokenized text better. We must explicitly split the question text string into different tokens (words) with delimiters such as space.

- **POS tagging and lemmatization**: Words can have different forms and tenses. We want to keep the words that are informative for our analysis while filtering out others. We use POS tagging to achieve this. POS tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context — i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. POS tagging can be achieved using the python TextBlob module. In the lemmatization the words are converted to its base form using the tags generated by TextBlob.

Word	**Lemmatization**
visualization	visualize
processing	process
words	word

table(3.1.1): lemmatization example

- **Vectorization**: It's difficult to work with text data while building Machine learning models since these models need well-defined numerical data. The process to convert text data into numerical data/vector, is called vectorization or in the NLP world, word embedding**.**
- **TF-IDF**: Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

a) **TF**: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more often in long documents than shorter ones. Thus, the term frequency is often

divided by the document length (aka. the total number of terms in the document) as a way of normalization:

$$\mathrm{TF(t,d)} = \frac{\text{number of times t appears in d}}{\text{total number of words in d}}$$

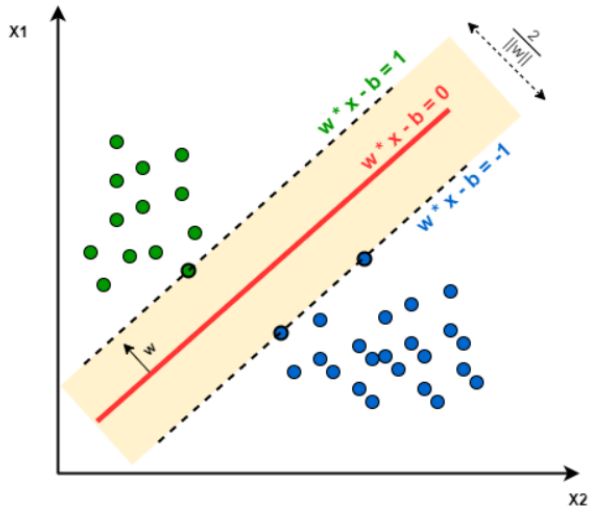
b) **IDF**: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. It's a logarithmic transformation of a fraction, calculated as the total number of documents in the corpus divided by the number of documents containing the word. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$\mathrm{IDF(t)} = \log \frac{\text{Total number of documents}}{\text{Number of documents that contain t}}$$

Training the dataset:

After converting the text data to numerical data, It's time to train a machine learning model on the vectorized dataset and test it.

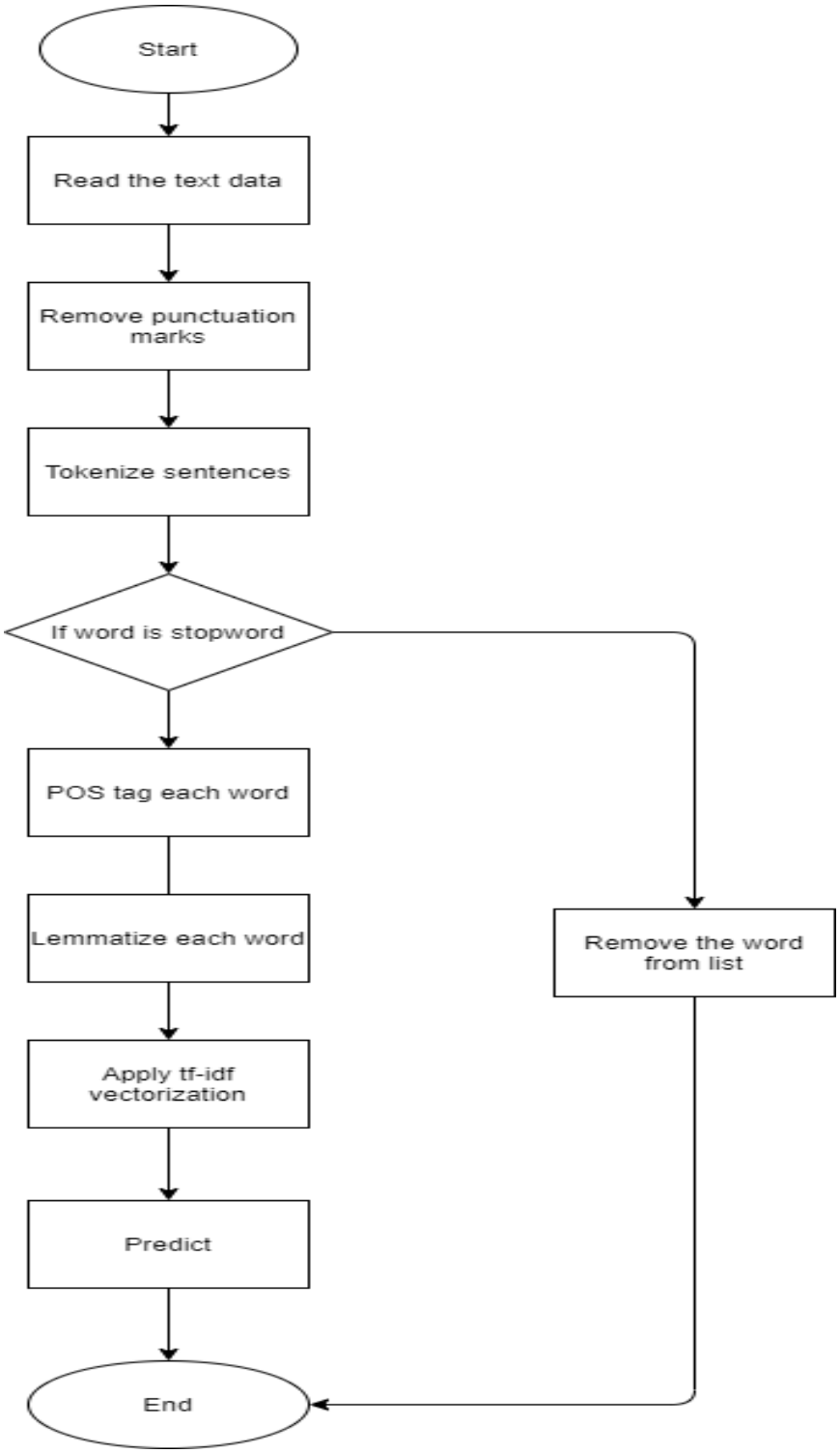
- **SVM**: Support Vector Machine (SVM) is a supervised machine learning algorithm. SVM's purpose is to predict the classification of a query sample by relying on labeled input data which are separated into two group classes by using a margin. Specifically, the data is transformed into a higher dimension, and a support vector classifier is used as a threshold (or hyperplane) to separate the two classes with minimum error. In the base form, linear separation, SVM tries to find a line that maximizes the separation between a two-class data set of 2-dimensional space points. To generalize, the objective is to find a hyperplane that maximizes the separation of the data points to their potential classes in an n-dimensional space. The data points with the minimum distance to the hyperplane (closest points) are called *Support Vectors*. In the image below, the Support Vectors are the 3 points (2 blue and 1 green) laying on the scattered lines, and the separation hyperplane is the solid red line:



fig(3.1.2): support vectors

- **LinearSVC**: The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function. The kernel method can not be changed in linear SVC, because it is based on the kernel linear method.
- **Multi-Class Classification using SVM**: Classification means categorizing data and forming groups based on the similarities. In a dataset, the independent variables or features play a vital role in classifying our data. When we talk about multiclass classification, we have more than two classes in our dependent or target variable. As the classification is taking place between 100 labels then multiclass classification is to be used. OneVsRest approach is used for classification where the breakdown is set to a binary classifier per each class. A single SVM does binary classification and can differentiate between two classes. So that, according to the breakdown approach, to classify data points from m classes data set the classifier can use m(m+1)/2 svms

Data flow for prediction:



fig(3.1.3): data flow for prediction

Examples:

Sentence	Tokenization	Removing stopwords and punctuations	POS Tagging	Lemmatization	Predicted class
'using angular for frontend, spring for backend'	['using', 'angular', 'for', 'frontend', ',', 'spring', 'for', 'backend']	['using', 'angular', 'frontend', 'spring', 'backend']	[('using', 'v'), ('angular', 'a'), ('frontend', 'n'), ('spring', 'n'), ('backend', 'n')]	['use', 'angular', 'frontend', 'spring', 'backend']	angularjs spring
'we used python for data science machine learning'	['we', 'used', 'python', 'for', 'data', 'science', 'machine', 'learning']	['used', 'python', 'data', 'science', 'machine', 'learning']	[('used', 'v'), ('python', 'n'), ('data', 'n'), ('science', 'n'), ('machine', 'n'), ('learning', 'n')]	['use', 'python', 'data', 'science', 'machine', 'learning']	python
'JavaScript and Java have been multipurpose programming languages '	['JavaScript', 'and', 'Java', 'have', 'been', 'multipurpose', 'programming', 'languages']	['javascript', 'java', 'multipurpose', 'programming', 'languages']	[('javascript', 'n'), ('java', 'n'), ('multipurpose', 'v'), ('programming', 'a'), ('languages', 'n')]	['javascript', 'java', 'multipurpose', 'programming', 'language']	java javascript

table(3.1.2): examples of data pre-processing and prediction

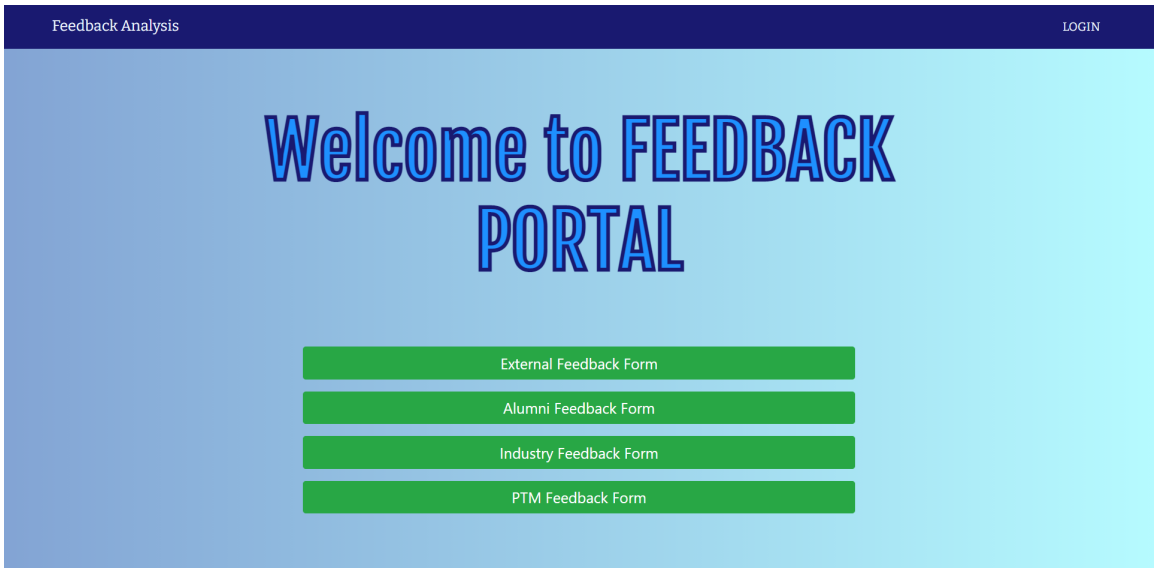
3.2. Hardware/Software Requirement

- **Hardware:**
  - Intel i3 processor
  - 4 GB RAM
  - 500 GB Hard disk
- **Software:**
  - Microsoft Windows 7+ or Linux.
  - Python version 3.
  - Python Libraries and modules:
    - Django
    - NLTK
    - Sklearn
    - TextBlob
    - Pickle
    - Pandas
    - Numpy

Chapter 04

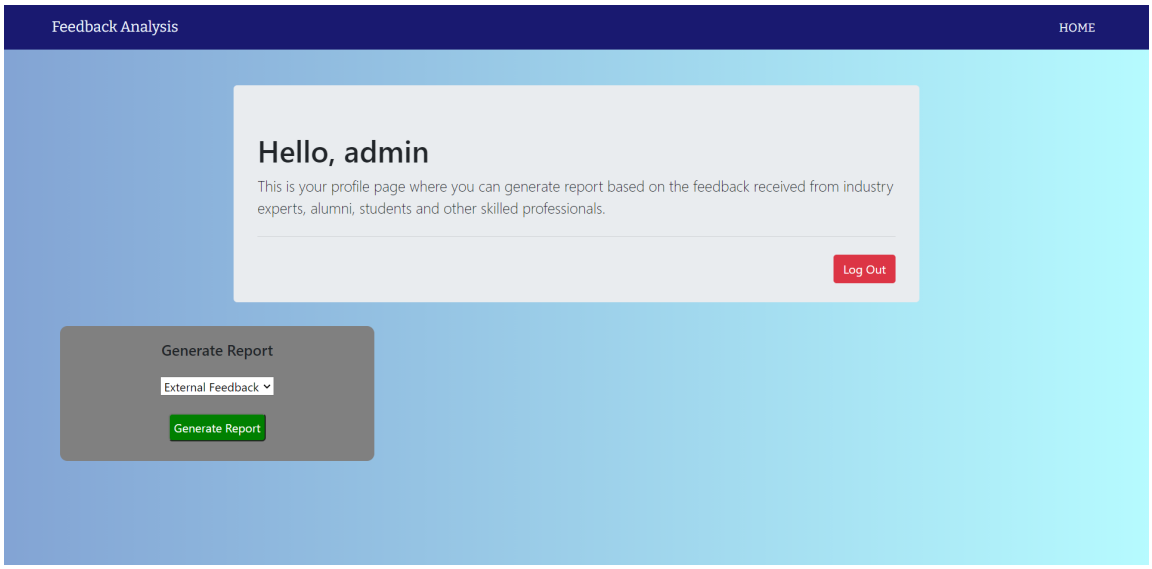
4.1. Project Results :

- Front Page:



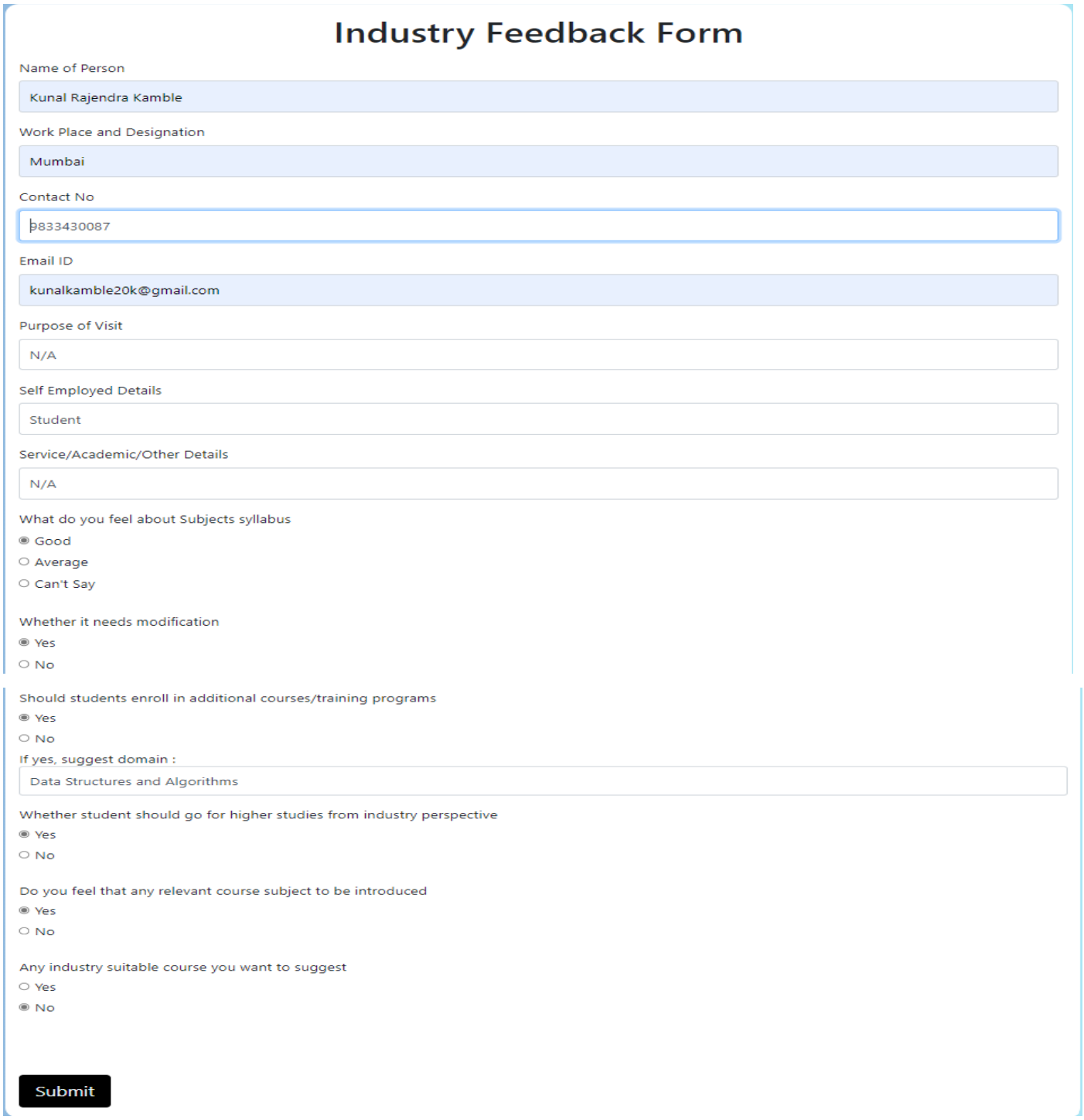
fig(4.1.1): front page

- Admin Interface:



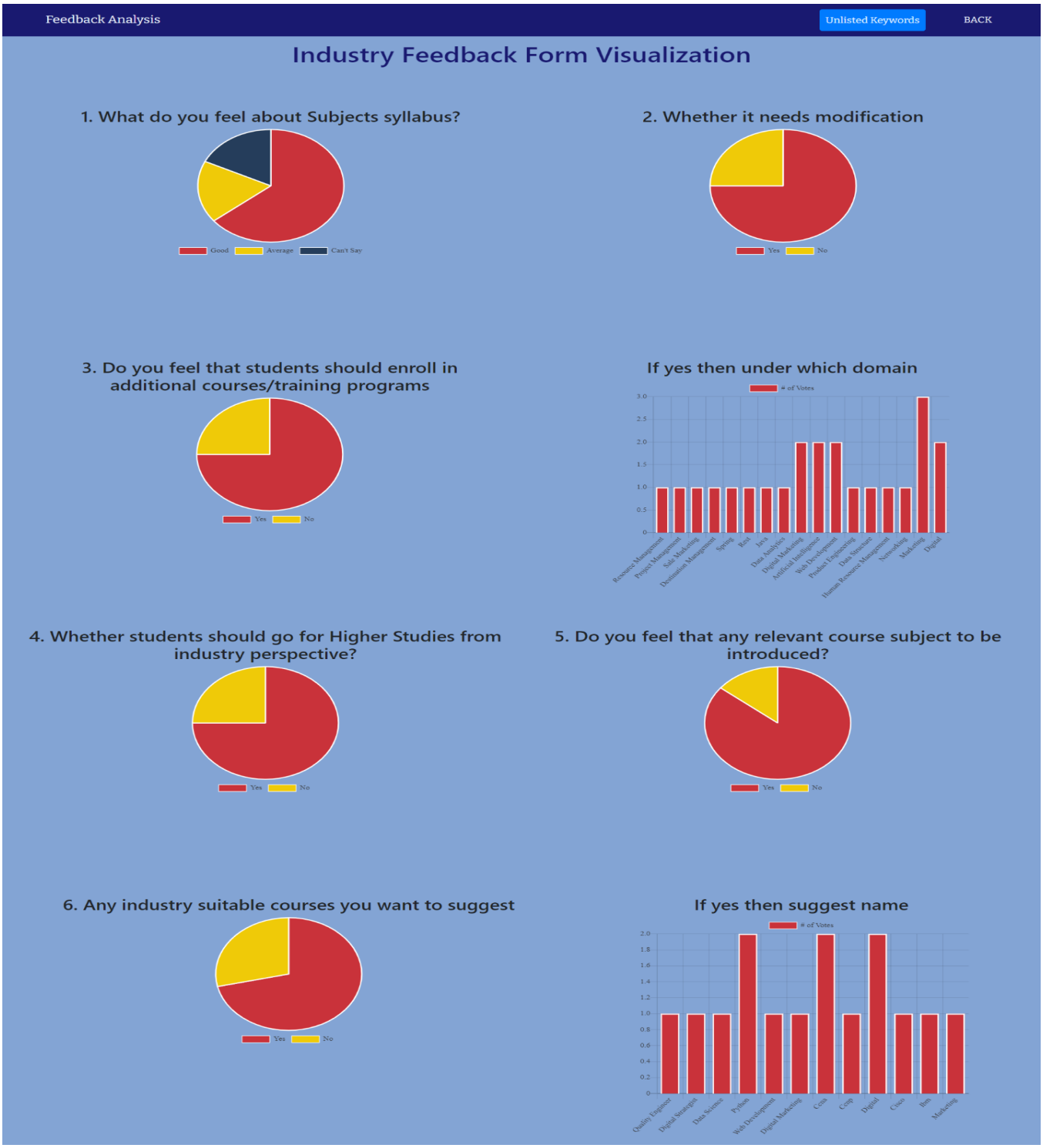
fig(4.1.2): admin interface

- Industry Feedback form:



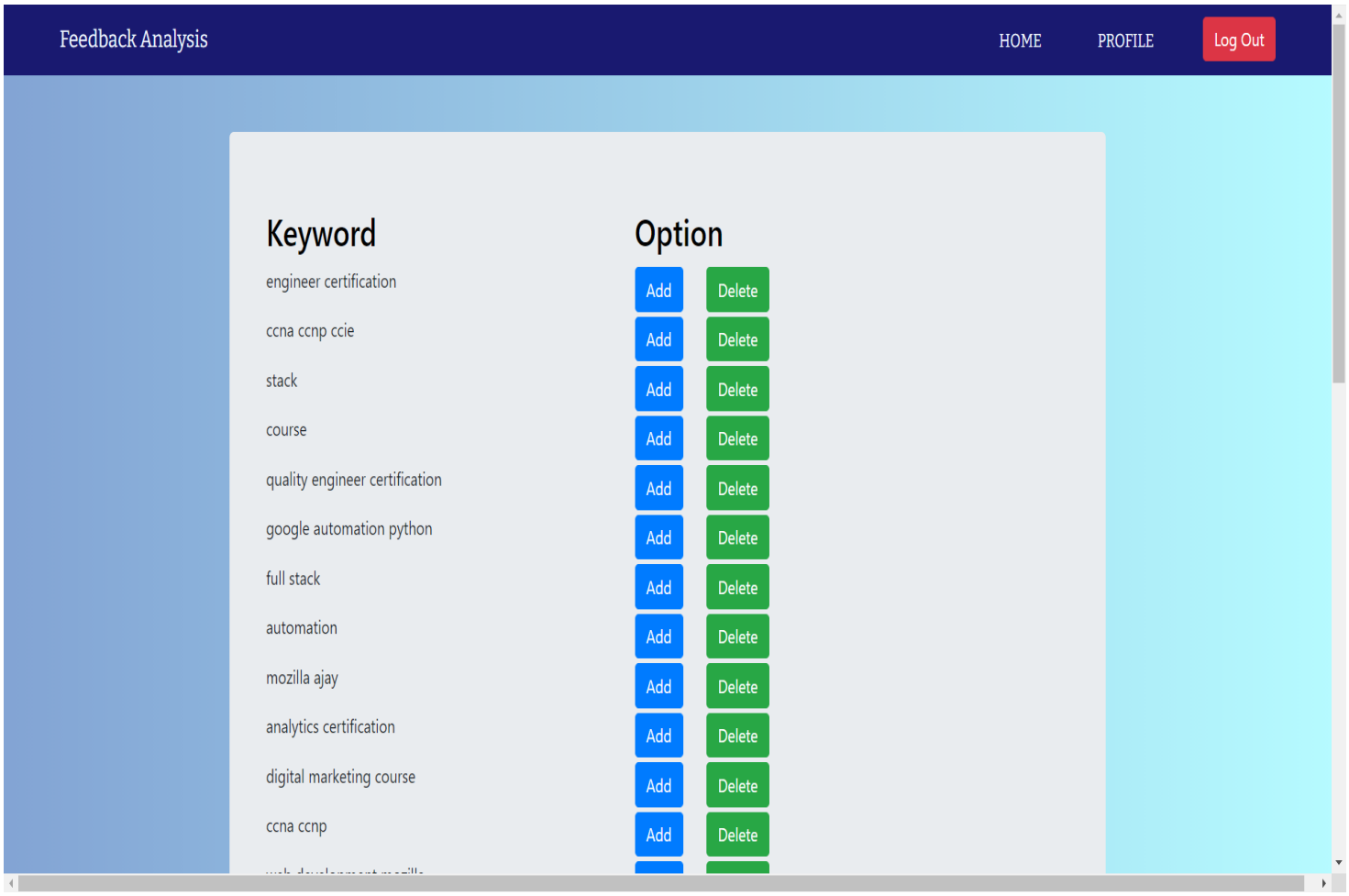
fig(4.1.3): industry feedback form

- Analysis and Visualization of the Industry Feedback form:



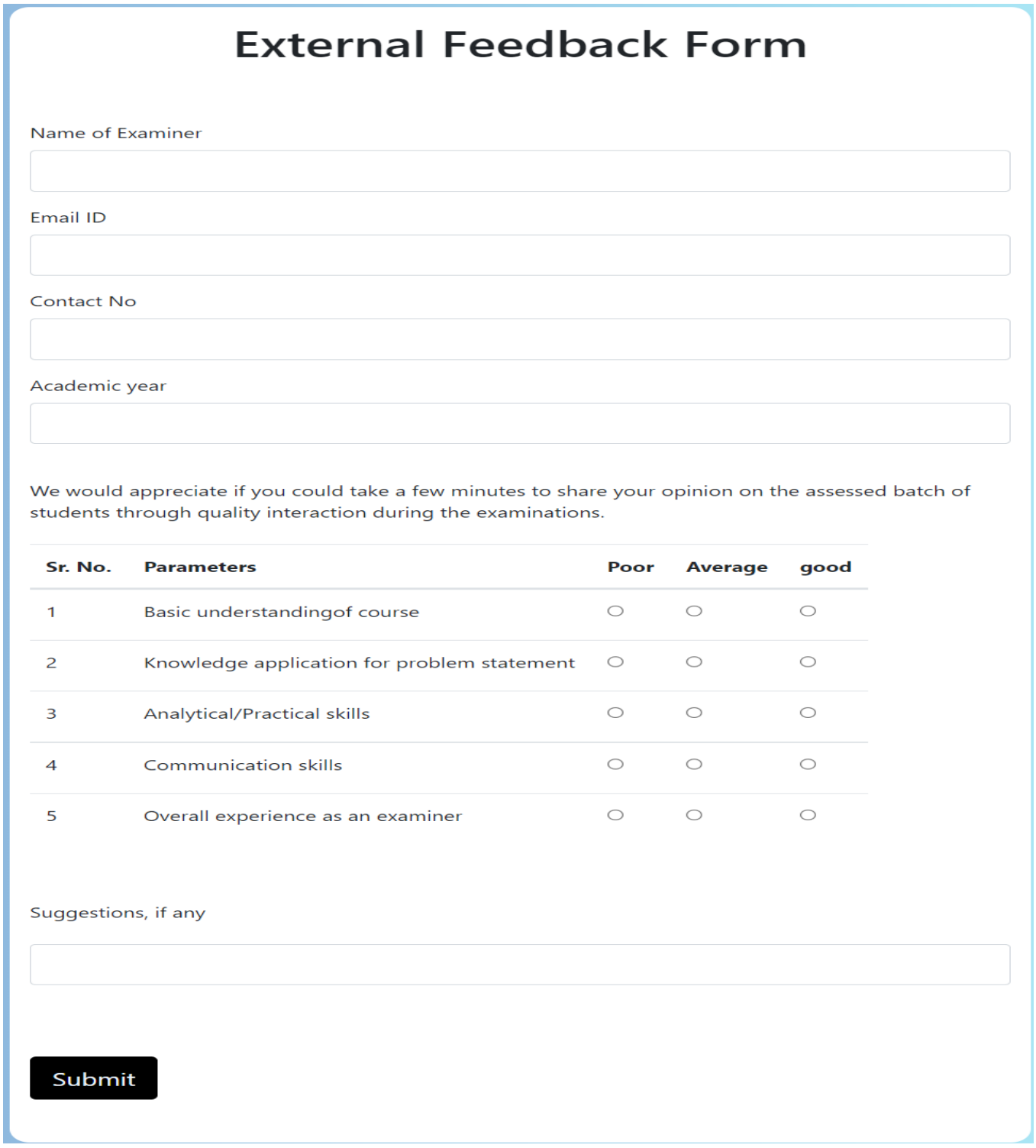
fig(4.1.4): industry feedback form visualization

- Unlisted keywords:



fig(4.1.5): unlisted keywords

- External Feedback Form:



External Feedback Form

Name of Examiner

Email ID

Contact No

Academic year

We would appreciate if you could take a few minutes to share your opinion on the assessed batch of students through quality interaction during the examinations.

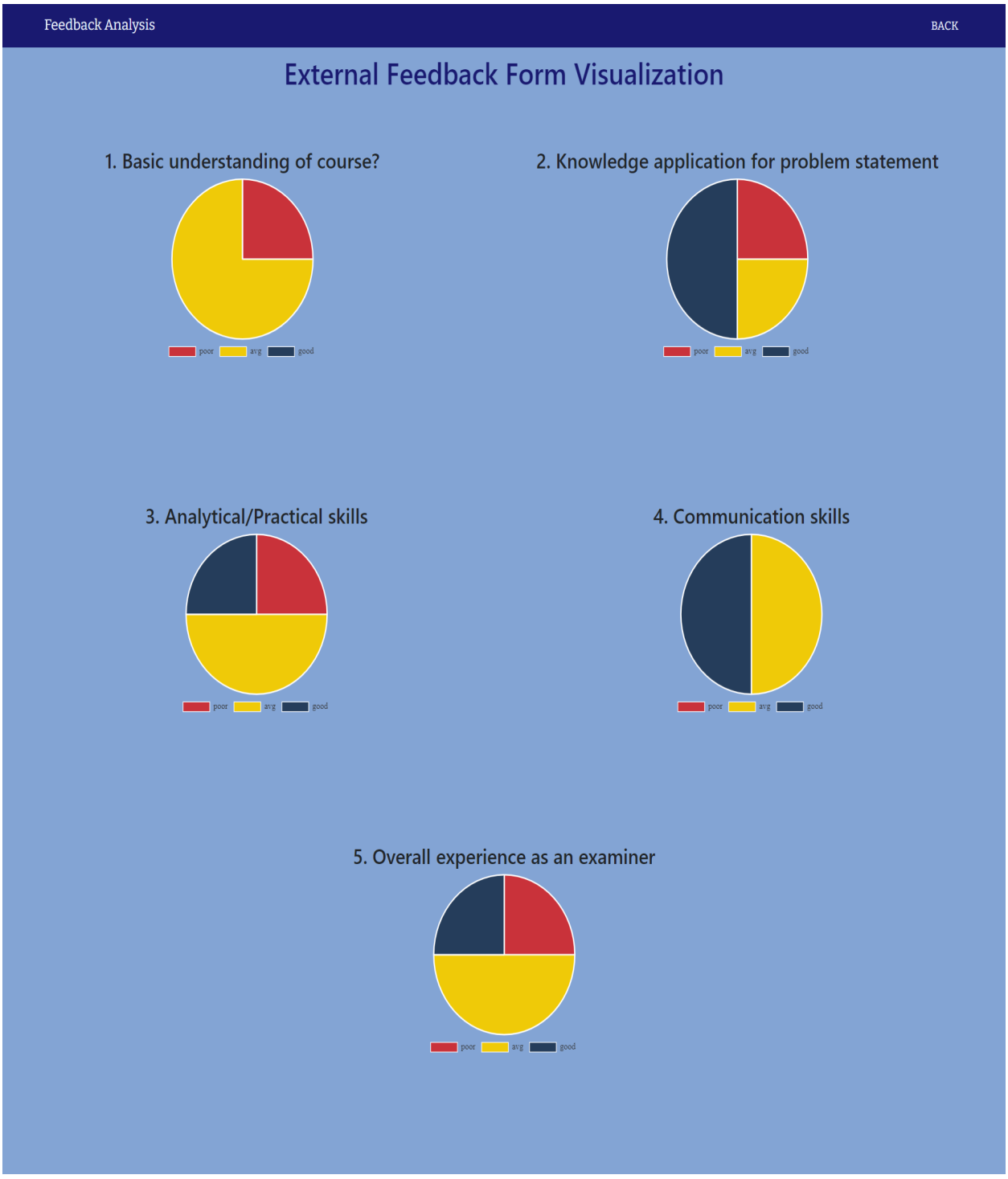
Sr. No.	Parameters	Poor	Average	good
1	Basic understandingof course	○	○	○
2	Knowledge application for problem statement	○	○	○
3	Analytical/Practical skills	○	○	○
4	Communication skills	○	○	○
5	Overall experience as an examiner	○	○	○

Suggestions, if any

Submit

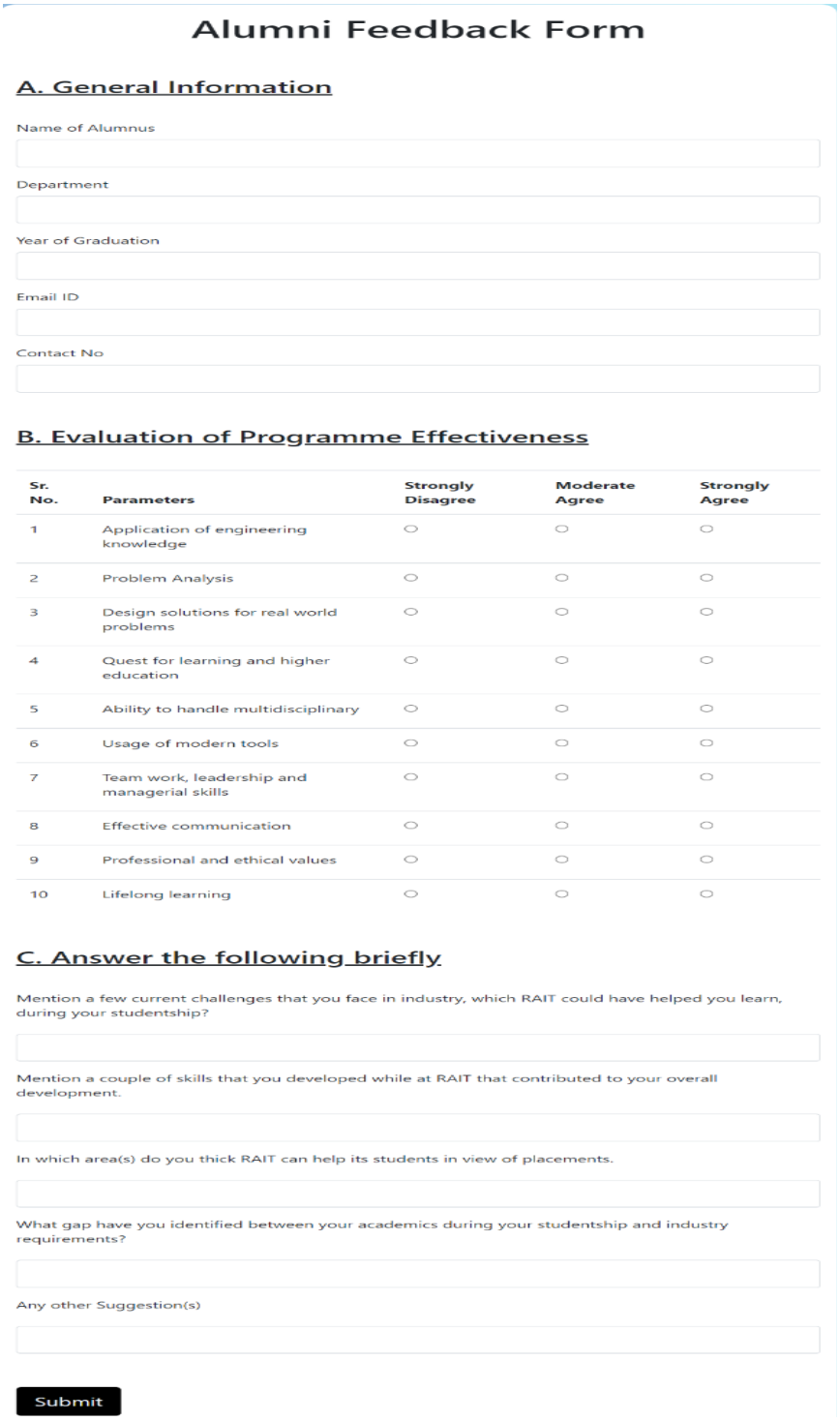
fig(4.1.6): external feedback

- Analysis and Visualization of the External Feedback form:



fig(4.1.7): external feedback form visualization

- Alumni Feedback Form:



Alumni Feedback Form

A. General Information

Name of Alumnus

Department

Year of Graduation

Email ID

Contact No

B. Evaluation of Programme Effectiveness

Sr. No.	Parameters	Strongly Disagree	Moderate Agree	Strongly Agree
1	Application of engineering knowledge	○	○	○
2	Problem Analysis	○	○	○
3	Design solutions for real world problems	○	○	○
4	Quest for learning and higher education	○	○	○
5	Ability to handle multidisciplinary	○	○	○
6	Usage of modern tools	○	○	○
7	Team work, leadership and managerial skills	○	○	○
8	Effective communication	○	○	○
9	Professional and ethical values	○	○	○
10	Lifelong learning	○	○	○

C. Answer the following briefly

Mention a few current challenges that you face in industry, which RAIT could have helped you learn, during your studentship?

Mention a couple of skills that you developed while at RAIT that contributed to your overall development.

In which area(s) do you thick RAIT can help its students in view of placements.

What gap have you identified between your academics during your studentship and industry requirements?

Any other Suggestion(s)

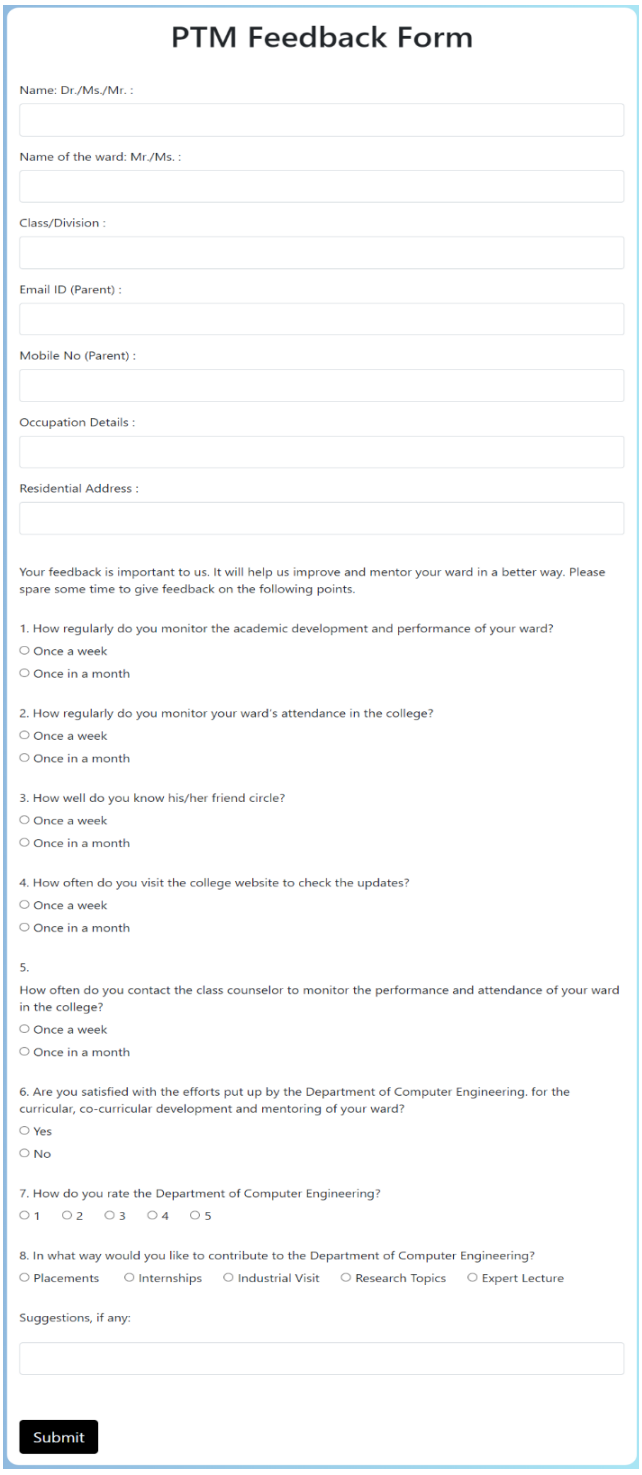
Submit

fig(4.1.8): alumni feedback form

- Analysis and visualization of the Alumni feedback form:

fig(4.1.9): alumni feedback visualization

- PTM Feedback Form:



PTM Feedback Form

Name: Dr./Ms./Mr. :

Name of the ward: Mr./Ms. :

Class/Division :

Email ID (Parent) :

Mobile No (Parent) :

Occupation Details :

Residential Address :

Your feedback is important to us. It will help us improve and mentor your ward in a better way. Please spare some time to give feedback on the following points.

1. How regularly do you monitor the academic development and performance of your ward?
○ Once a week
○ Once in a month

2. How regularly do you monitor your ward's attendance in the college?
○ Once a week
○ Once in a month

3. How well do you know his/her friend circle?
○ Once a week
○ Once in a month

4. How often do you visit the college website to check the updates?
○ Once a week
○ Once in a month

5.
How often do you contact the class counselor to monitor the performance and attendance of your ward in the college?
○ Once a week
○ Once in a month

6. Are you satisfied with the efforts put up by the Department of Computer Engineering. for the curricular, co-curricular development and mentoring of your ward?
○ Yes
○ No

7. How do you rate the Department of Computer Engineering?
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

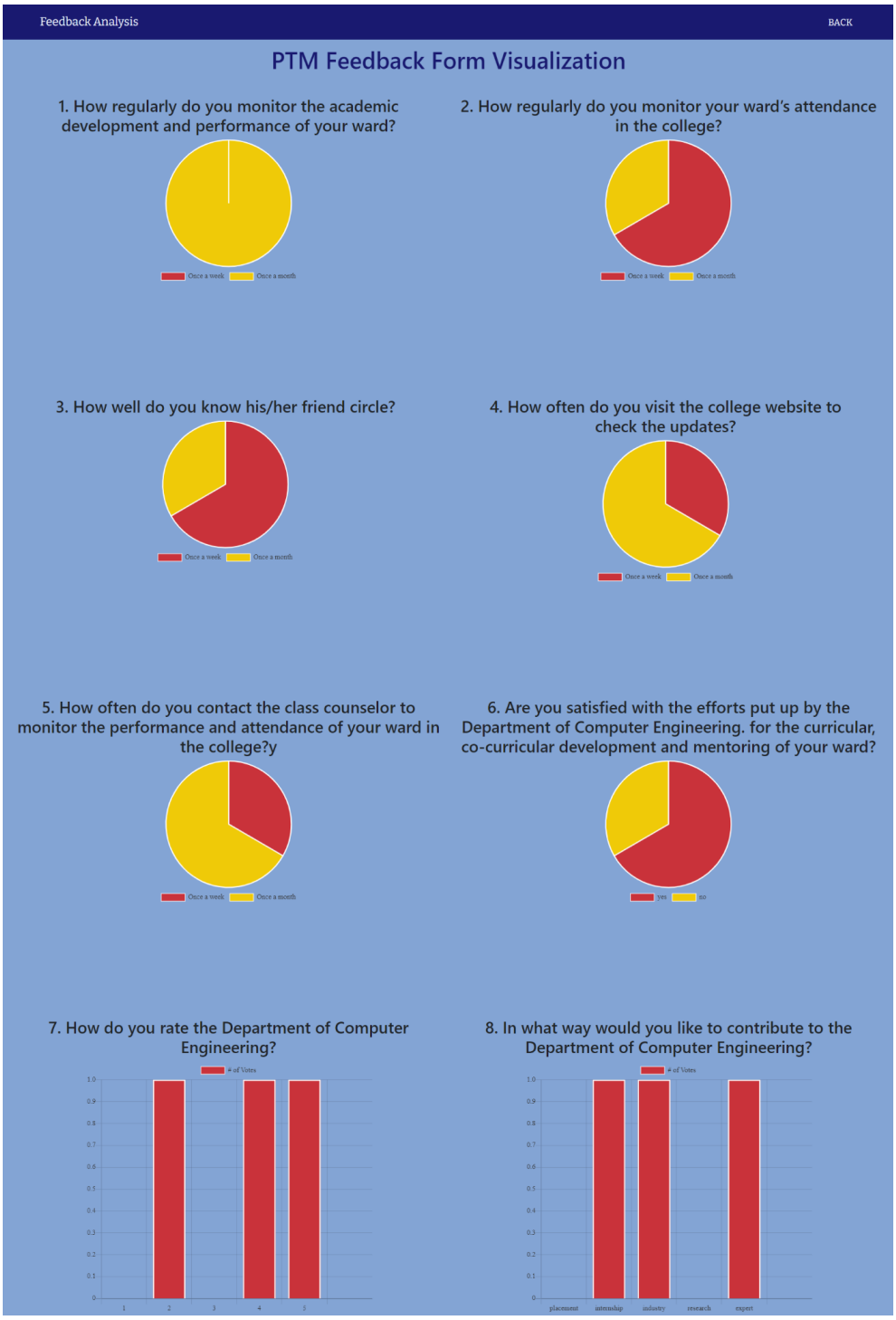
8. In what way would you like to contribute to the Department of Computer Engineering?
○ Placements ○ Internships ○ Industrial Visit ○ Research Topics ○ Expert Lecture

Suggestions, if any:

Submit

fig(4.1.10): ptm feedback form

- Analysis and visualization of the PTM feedback form:



fig(4.1.11): ptm form visualization

Chapter 05

Benefits of Proposed Work :

- The proposed system not only performs analysis and visualization on choice based data but also analyses textual data.
- The NLP model provides classification for 100 computer related terminologies, which is important for analysis of textual data.
- The NLP model can recognize 100 labels with an accuracy of 76%.
- Since there are more terminologies getting introduced in the field of computer science, the analysis algorithms will get updated to recognize more terminologies with the admin interface.
- The admin can decide which unrecognized terminology should be considered for further analysis and which should be dropped from the analysis process.

Chapter 06

Conclusion :

External Feedback Analysis ,NLP based project that helps the organization to get a detailed analysis of forms filled by the Alumni, External Examiner, Industrymen and Parents of Students regarding Parent Teachers meeting.

Although several notable works have come in this field, a fully automated and highly efficient system has not been introduced till now. This is because of the unstructured nature of natural language. The vocabulary of natural language is so large that things become even harder. Several challenges still exist in the field of machine learning . These problems have to be tackled separately and those solutions can be used to improve the methods to do analysis.

The proposed machine learning model can recognize 100 computer related terminologies with the accuracy of 76%. As the ml model cannot recognize every computer related terminology, the terms which are not recognized by the ml model are stored in the database for the admin to either consider it for further analysis or reject it which is functional to admin only.

Chapter 07

References :

[1] The Application of NLTK Library for Python Natural Language Processing in Corpus Research, Meng Wang, Fanghui Hu, Sep-2021.
[2] Performance Comparison of Crowdworkers and NLP Tools on Named-Entity Recognition and Sentiment Analysis of Political Tweets, Mona Jalal, Kate K. Mays, Lei Guo, and Margrit Betke, Feb-2020.
[3] SARCASM AS A CONTRADICTION BETWEEN A TWEET AND ITS TEMPORAL FACTS: A PATTERN BASED APPROACH, Santosh Kumar Bharti and Korra Sathya Babu, Oct-2018
[4] INFORMATION RETRIEVAL TECHNIQUE FOR WEB USING NLP Rini John and Sharvari Govilkar, Oct-2017
[5] NATURAL LANGUAGE PROCESSING USING PYTHON, Vismaya V, Darvin Reynald J, May-2017.

Chapter 08

Acknowledgement :

We would like to express our special thanks of gratitude towards the Chairman, the Vice Chairman and the members of software Development Group, RAIT for providing us with an amazing opportunity to work on this project through the SDG Internship program.
We extend our sincere and most heartfelt thanks towards Dr. Leena Ragha Mam (HOD of Computer Engineering Department) and Mrs. Saguna Ingle Mam, for providing us guidance throughout the project.
All the support and encouragement of the aforementioned has been an important factor towards the completion of the project.