MUHAMMAD MOAVIA
01-134182-072
SHAMIR JAMAL ABBASI
01-134182-056

Expert Ranking in Social Question Answering Forum

Bachelor of Science in Computer Science

Supervisor: Dr. Saba Mahmood

Department of Computer Science
Bahria University, Islamabad

June, 2022

Certificate

We accept the work contained in the report titled "EXPERT RANKING IN SOCIAL QUESTION-ANSWERING FORUMS", written by MUHAMMAD MOAVIA AND SHAMIR JAMAL ABBASI as a confirmation to the required standard for the partial fulfillment of the degree of Bachelor of Science in Computer Science.

Approved by . . . :

Supervisor: Dr. Saba Mahmood (Assistant Professor)

_____

Internal Examiner:

_____

External Examiner:

_____

Project Coordinator: Dr. Moazzam Ali (Assistant Professor)

_____

Head of the Department: Dr. Arif ur Rahman (HOD/ Sr. Associate Professor)

_____

June 22^{nd}, 2022

Abstract

The CQA(Community question answering) is a perfect platform for the people who frequently participates to get the desired information of their interests. But if we want to find the expertise kind of users with relevant and authentic answers is the key challenge within these communities. In case, if we have to trust on someone's opinion who is not well known by the communities, it is mandatory to find the credibility of the user. Because there is a scarcity of specialist organisations, professionals must rely on other resources when seeking knowledge. A huge online community, such as a Facebook discussion group, may have millions of members with a vast knowledge base containing millions of text documents. But a poor or low quality answer shows unqualified users therefore a priority is to find expert users. Expert-finding systems now in use assess user expertise based on the content of produced papers or one's social position. We developed an expert finding technique called ExpertRank in this study, which assesses a user's knowledge based on the content of authored documents and the social link of the authors. To identify the importance of topic, we used Latent Dirichlet Allocation (LDA) model to evaluate textual documents. Then we identified the social importance of users by using Google's PageRank algorithm which is used by Google to rank websites. The modified PageRank calculates the social importance of authors who answered on same topic at different places. The more the social importance of authors, the higher the chances of being an expert. After that, we combine these expert ranking techniques using the cascade combine strategy. We evaluate our proposed algorithm ExpertRank using a most popular online community platform StackOverflow. The experiments show that the proposed algorithm performs the best operations and shows the best performance when both topic modeling and social link analysis are considered.

Acknowledgments

To start off we would like to thank our supervisor Dr. Saba Mahmood. We acknowledge Dr. Saba Mahmood remarkable guidance regarding academic as well as administrative procedures. Being supervisor for this project, she managed to provide us with her valuable feedback and suggestions on our project. She was always there as a mentor for us whenever we needed her supervision. We are grateful to Dr. Saba Mahmood for her cooperation support in evaluation of our research and providing us with her valuable feedback throughout the research work. We would like to thank here to our family who made all this project possible. It could not be possible without their support in any regard. Our sincere thanks to Head of Computer Science Department at Bahria University Islamabad for providing us with a great opportunity to conduct our research at the institution and also for providing us access to the resources and researches of this institution.

MUHAMMAD MOAVIA
Bahria University Islamabad, Pakistan

SHAMIR JAMAL ABBASI
Bahria University Islamabad, Pakistan

June 22^{nd}, 2022

"Failure isn't a problem. It's the fear of failure that's the limiting factor. You can't lose your nerve for the big failure, because it's the exact same nerve you need for the big success."

Regina Dugan

Contents

Abstract i

1 Introduction 1
- 1.1 Project Background/Overview . . . 1
- 1.2 Problem Description . . . 2
- 1.3 Objectives . . . 2
- 1.4 Project Scope . . . 2
- 1.5 Feasibility Study . . . 3
  - 1.5.1 Risks Involved . . . 3
  - 1.5.2 Resource Requirement . . . 3
- 1.6 Solution Application Areas . . . 3

2 Literature Review 5

3 Requirement Specifications 12
- 3.1 Existing System . . . 12
- 3.2 Proposed System . . . 12
- 3.3 Requirement Specifications . . . 13
  - 3.3.1 Functional Requirements . . . 13
  - 3.3.2 Non-functional Requirements . . . 15
  - 3.3.3 Domain Requirements . . . 15
- 3.4 Use Case Diagram/ Descriptive Use Case . . . 17
- 3.5 System Use Case . . . 19
  - 3.5.1 Post Question/ Add Description: . . . 21
  - 3.5.2 Responses . . . 23
  - 3.5.3 Find Appropriate Answer: . . . 25

4 Design 27
- 4.1 System Architecture . . . 27
- 4.2 Design Constraints . . . 29
  - 4.2.1 Programming Language/Techniques . . . 29
  - 4.2.2 Resources Used . . . 30
  - 4.2.3 Assumptions and Dependencies . . . 30
- 4.3 Design Methodology . . . 30
- 4.4 High Level Design . . . 32
  - 4.4.1 Sequence Diagram . . . 32
  - 4.4.2 Context Diagram . . . 36

4.4.3 Class Diagram . . . 37
4.5 Low Level Design . . . 39
4.5.1 Block Diagram . . . 39
4.6 Graphical User Interface Design . . . 40
4.6.1 Select Topic . . . 41
4.6.2 Topic Details . . . 41
4.6.3 Expert Author/s . . . 41
4.6.4 Visualizing Experts . . . 42
4.7 External Interfaces . . . 43

5 System Implementation **44**
5.1 System Architecture . . . 44
5.1.1 Data Pre-processing . . . 44
5.1.2 Applying Natural Language Processing (NLP) Techniques: . . . 45
5.1.3 Text Modeling . . . 45
5.1.4 Link Analysis . . . 47
5.2 Tools and Technologies . . . 50
5.2.1 Libraries . . . 50
5.2.2 Frameworks/ Languages . . . 52

6 System Testing and Evaluation **53**
6.1 Software Testing Techniques . . . 53
6.2 Graphical User Interface Testing . . . 54
6.3 Usability Testing . . . 54
6.4 Unit Testing . . . 54
6.5 System Testing . . . 54
6.6 Black Box Testing . . . 55
6.7 White Box Testing . . . 55
6.8 Acceptance Testing . . . 55
6.9 Test Case . . . 56
6.9.1 Test Case 1: Validating the Existence of Questioner . . . 56
6.9.2 Test Case 2: Validating the Presence of Author . . . 57
6.9.3 Test Case 3: Topic Modeling . . . 58
6.9.4 Test Case 4: Verifying the Links of Authors . . . 59
6.9.5 Test Case 5: Finding the Experts . . . 60

7 Conclusions **61**

A Data Dictionary **62**

References **65**

List of Figures

3.1 Use-case Dataset Scrapping 17
3.2 Use-case Diagram 19
3.3 Use-case PostQuestion 21
3.4 Use-case Responses 23
3.5 Use-case Ranking 25

4.1 Architecture Diagram 28
4.2 Flowchart Diagram 31
4.3 Sequence Diagram 32
4.4 Scrapping Sequence Diagram 33
4.5 Question Sequence Diagram 33
4.6 Answers Sequence Diagram 34
4.7 Ranks Sequence Diagram 35
4.8 Context Diagram 36
4.9 Class Diagram 38
4.10 Block Diagram 40
4.11 Topic Selection 41
4.12 Topic Details 41
4.13 Authors 42
4.14 Topic Experts 42

5.1 Word Cloud Diagram 46
5.2 Top Most Salient Terms 47
5.3 Topic Author Relationship Graph 48
5.4 Total Responses On Particular Topic 49

List of Tables

2.1 Research analysis 5
2.2 Research analysis 6
2.3 Research analysis 7
2.4 Research analysis 7
2.5 Research analysis 8
2.6 Research analysis 9
2.7 Research analysis 9
2.8 Research analysis 10
2.9 Research analysis 11

3.1 System Scrapping Use-Case 18
3.2 System Use-Case 20
3.3 Use-Case PostQuestion 22
3.4 Use-Case Manage Responses 24
3.5 Use-Case Appropriate Answers 26

6.1 Test Case Validating Questioner 56
6.2 Test Case Validating Expert 57
6.3 Test Case Topic Modeling 58
6.4 Test Case Link Analysis 59
6.5 Test Case Finding Experts 60

A.1 Data Dictionary 62

Acronyms and Abbreviations

CQA	Community Question Answering
NLP	Natural Language Processing
ERS	Expert Recommender System
PR	Page Rank
LDA	Latent Dirichlet Allocation
SNA	Social Network Analysis
API	Application Programming Interface
LA	Link Analysis
TM	Topic Modelling

Chapter 1

Introduction

1.1 Project Background/Overview

Community Question Answering (CQA) forums allow users to ask questions of their interests and get answers from the users who are interested in the relevant topic. This is done for providing social support and it attracted users in various fields. Many questions are posted on these sites every day. Community question answering is just a framework like search engines that retrieves information against available data. There are many online communities question answering sites like Quora, StackOverflow, Microsoft discussion group, etc. where a lot of people use these frameworks to ask their questions. Generally, answers are provided by the experts according to their understanding of the topic relevant to the question. The question answering content in online community forums benefits the users to obtain information in the form of answers by other users.
The problem of community question answering forum is the lack of topic expert of the relevant question. Many users do not understand the questions, or they do not have prior knowledge of the relevant topic and still, they answer the questions. So, it becomes difficult for the user who posted the query to find out which person has provided the right answer and if that person is the expert on a relevant topic or not. It interprets and presents data by Natural Language Processing (NLP) techniques. The challenging task is to rank the expert based on the profile of the expert and the quality of the answer. So, we will propose an expert ranking algorithm for discussion groups based on online communities which are called Expert Rank where we will use expert profiles built from discussion groups to rank the authors of the answers.

1.2 Problem Description

Since the enhancement in web 2.0 has increased the availability and popularity of systems based on user-generated content. Community question answering websites or forums such as Quora, StackOverflow, Microsoft discussion forums are leading for the past few years. A community question answering forum may have tens of thousands of questions posted every day. The growing number of new questions and bulk replies to those questions makes it more difficult for a general user to find the appropriate answerer to the question and the answer itself. There is an increasing failure rate because it becomes difficult to rank the expert of question. Some current community question answering forums tell the upvotes of a good answer. But we don't know the person who has provided the answer is a topic expert to relevant answer or not. Usually, an answer is upvoted but technically and logically it is not correct. Some of the answers are which are not initially written and at the bottom are not read and considered by many users. To resolve the problems proposed above, the ExpertRank system ranks to potential answerers who are most likely to provide satisfying answers. We will develop an ExpertRank system where users posted different questions according to their interests and the questions are answered by many authors. The authors will be identified according to their degree of relevance to the question, some of the features based on the author's profile. The quality of the answer will also be identified from the user's text.

1.3 Objectives

- To develop a system to rank/find the expert of the query posted by a user based on expert's profile information, the quality of answer usually referred to as expert finding.
- To develop such an algorithm whose methodology and techniques can be used in different question-answering forums to help the community.
- To make our system commercially viable by integrating it with different discussion forums.

1.4 Project Scope

The Expert ranking system is used to portray the benefit and efficiency of online community discussion forums or blogs. Traditional question-answering forums can give the quality answer but if we get to know that the answer had been provided by a relevant topic expert then this will make the answer authorized. Hence, expert ranking in the community

question answering forum provides high-quality and useful answers. Therefore, the expert should be familiar with the topic of the question asked and provide the answer clearly.

1.5 Feasibility Study

The community question-answering forums should be able enough to provide actual information which is relevant to the problem. When such a system with required modifications will be implemented, it will enhance question-answering in the community. The criteria to evaluate the answer is to implement link and content analysis techniques as described above in the system. The solution provided by using these techniques will be more feasible than the other systems.

1.5.1 Risks Involved

There are some of the risks which are involved in our system. These risks will be tackled to avoid any inconvenience for implementing the system. The system should not take excessive time as this can lead to incomplete implementation. When the system is deployed after applying all the techniques, it can lead to unexpected behavior. Data overfitting is also a risk that can occur in the system. The system can show results with noise. So, overfitting must be removed from the system. The data compliance issue can be the problem. The information we extract can be biased which is not true according to our requirements. It is important that the model can be interpreted.

1.5.2 Resource Requirement

Getting the desired dataset is our first and foremost requirement. To implement the system, many software and hardware requirements should be met. The machine should have a solid-state drive for enhancing the performance of the system.

1.6 Solution Application Areas

Finding or ranking a topic expert in online community question answering is the need of every person. The current community question answering forums only provide answers that are also not reliable. The best answer can also be ignored. We don't know whether the answer is given by the expert author or a spam one. Nowadays when everyone is familiar with Internet browsing and using the internet and social media. When they see any news on any social media platform they rely on that news without any validation or authentication and did not know whether the news is true or not. This system enables the users to find their quality answers which are given by the experts. This system can be integrated with question-answering blogs or community discussion forums. The systems can be installed

on mobile applications where people will post the question and get replies from the authors. The system algorithms and techniques will be working on the background and the perfect answer related to the problem asked can be marked by the application. Similarly, Facebook groups can use this system to rank which author has provided a suitable answer to the problem.

Chapter 2

Literature Review

Discussion groups and forums, a new type of web-based community which allow users to share their knowledge and experience and provide social support, that attract many users in different disciplines. A new expert ranking algorithm for discussion groups based on the online community has been proposed as expert ranking. Expert ranking uses both domain-driven and domain-independent information and uses a modified PageRank algorithm to calculate expert authorization ratings. We then compute the authority scores and relevance scores, respectively.

Table 2.1: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
J.Jiao, J. Yan, H. Zhao	ExpertRank: An Expert User Ranking Algorithm in Online Communities	6

The PageRank algorithm is comparable to the Google search algorithm. This algorithm is used to calculate the expert relevance score. The score is calculated after creating a directed edge from user to user. The weighted reference relationship algorithm distributes weights to topic starter and topic replier pairings and validates that the answer does not come from the individual who asked the question. The TFIDF gives each term in the document a weight based on how often it appears in the document. Words having a higher frequency are seen as more significant. The expert ranking system makes it possible to locate the most relevant experts for specific questions, and the authority earned via the

expert network dramatically improves expert discovery.

There are many beneficial knowledge services are provided by community Question Answering (CQA) site to Online user. Yahoo! Answers, Stackoverflow, Ubuntu, Wikipedia, and other online discussions Forums are prominent example of CQA services. Stackoverflow provides ability to search for posted questions, tags and users. Experts are found using link analysis techniques based on relationship between question and answers. This paper, expert ranking method using g-index is suggested and applies to stackoverflow forum records.

Table 2.2: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Husain	Expert finding systems: A systematic review.	10

Exp-PC, Rep-FS, and Weighted Exp-PC are three approaches presented. ExpPC is G-index adaptation for ranking experts in Stackoverflow forums. Experimental results of the proposed expert ranking method. Several features in Rep-FS, such as voter reputation and votes ratio, are offered to assess a user's skill. Exp-PC and Weighted Exp-PC, in particular, demonstrate that these approaches are more effective at identifying true experts.

This study proposes a new framework for expert mining in online communities. Thus, the proposed Expert Recommender System (ERS) builds a Trust-based ERS by using a well-known global-trust metric, PageRank, to find experts in a community question-answering (StackOverflow) and then using collaborative filtering to find similar experts based on their level of expertise and topics of interest to a specific user. In social network research, PageRank has been frequently utilized to identify experts in an online community.

Table 2.3: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Roy P.K., Jain A., Ahmad Z., Singh J.P	Identifying Expert Users on Question Answering Sites.	9

To begin, an API is used to retrieve a list of Top-k users based on their reputation points gained in the community from their website. As a result, we took data from the top-k users of the community question-answering system, applied our proposed framework to these users, and generated our own list of experts as an outcome. The purpose of this research is to see how effective the proposed system is. The suggested system ensures that an expert with appropriate field questions and interests is proposed.

With an influx of users and content on CQA platforms, the quality of their responses has recently sparked widespread concern. The user input on responses (i.e., the votes of answers) is used as the "relevance" labels in this study, which formalizes expert discovery as a learning to rank problems. The listwise learning to rank approach, also known as ListEF, is utilized to accomplish this goal. Recognizing that questions in CQA are typically brief and tagged, the ListEF technique proposes a tag word topic model (TTM) to generate high-quality topical representations of questions

Table 2.4: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
X. Cheng, S. Zhu, G. Chen and S. Su	Exploiting User Feedback for Expert Finding in Community Question Answering	4

A Competition-based User expertise Extraction (COUPE) approach for extracting user expertise features for given questions is created using the tag-word topic model (TTM). After gathering user expertise features, we utilize Lambda MART to train the ranking function using lists of users represented by feature vectors and their received votes as training instances. Finally, we can utilize the learned ranking algorithm to rank users for new queries and choose individuals with high ranks as prospective experts. Our project's

content analysis could be aided by the proposed system. This system's techniques can be used to perform content analysis.

This work introduces the Exp-rank algorithm, a novel expert ranking system. Exprank takes into account not only the authority of users in community but also the quality of the content that users upload.

Table 2.5: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Zhao, Nan Cheng, Jia Chen, Fei Cheng	A Novel Expert Finding System for Community Question Answering	5

We calculate the similarity between the new questions and the users' knowledge tags using expert rating. We recommend experts more appropriately to the new question based on the estimated results. The Exp-Rank takes into account users' ongoing performance and authority to provide a more impartial and comprehensive rating of experts. It also suggests experts based on the similarity between the new query and the expert users' knowledge tags. This Exp-rank algorithm can help in identifying similarities between question-and-answer relationships in our project.

Every day, tens of thousands of questions may be posted on a CQA website. Without enough collaboration assistance, the growing number of new questions could cause two difficulties for CQA systems. First, finding the right question to answer becomes more difficult for a general answerer. Furthermore, the quality of answers is unmanageable due to the unpredictability of the question-answering process, which involves a wide range of skill and education levels among answerers.

Table 2.6: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Zhengfa Yang, Baowen Sun	Expert recommendation in community question answering: a review and future direction	1

To make it easier for individuals who are just getting started with their study into CQA expert recommendation, and to show current trends and areas that need more attention from the research community.

Websites that allow people to contribute their expertise on open platforms, such as Community Question Answering (CQA), have grown into vast knowledge libraries. We introduced Topic Expertise Model (TEM), an unique probabilistic generative model with GMM hybrid, to simultaneously model topics and expertise by integrating textual content model and link structure analysis, to tackle this cluster of closely connected problems in a principled method.

Table 2.7: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Minghui Qiu, Swapna Gottipati, Liu Yang	CQARank: Jointly Model Topics and Expertise in Community Question Answering	11

We suggested CQARank to quantify user interests and knowledge score under different topics based on TEM results. Making "recommendations" for new questions is a key duty on CQA sites; the objective is to either lead questions to the correct expert users or answers, or to locate comparable questions for the asker to further explore related answers.

This research offers a thorough literature review to clarify the current state of the CQA literature that has employed ML and DL. The purpose is to summarise and consolidate the primary CQA research themes linked to I questions, (ii) answers, and (iii) users.

Table 2.8: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Dubey, Tondulkar	Analysis of community question-answering issues via machine learning and deep learning	3

ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, and Scopus were among the databases we used. These were found by searching for the terms 'CQA,' 'question answering,"social question answering,' 'expert users,' 'question quality,' and 'answer ranking' in the CQA literature. Most articles focused on a single platform, with only a few cross-platform inquiries. The number of articles with ML much outnumbers those with DL. DL's use in CQA research, on the other hand, is on the rise. A number of study avenues are suggested.

Using social media to locate expertise within an organisation is a frequent practice. People are not isolated, but are linked by a variety of ties. We suggest numerous ways for detecting people's associations from emails and web pages in our approach. We proposed an expertise propagation algorithm based on social networks: we select a small set of the top candidates as seed from a ranked list of candidates based on their probability of being an expert for a specific topic, and then use the social networks among the candidates to discover other potential experts.

Table 2.9: Research analysis

Author Name(s)	**Paper Title**	**Reference Number**
Yupeng Fu, Rongjing Xiang	Finding Experts Using Social Network Analysis	13

People explicitly send email to one another, therefore the relationship is likely to be encoded in the patterns of communication; on the other hand, we can rely on statistical correlations drawn from co-occurrences of people in web pages to build social networks.

Chapter 3

Requirement Specifications

3.1 Existing System

Many community question-answering websites or forums such as Quora, stack overflow, Microsoft discussion forums are led from the past few years. A CQA forum may have tens of thousands of questions posted every day. There are thousands or millions of people who post new questions and get a reply from different users on the daily basis. Random people come and post a question and get an answer from different random authors or experts. The growing number of new questions and bulk replies to those questions makes it become more difficult for a general user to find the appropriate answerer of the question and the answer itself.
Some current community question answering forums tells the upvotes of a good answer. But we don't know the person who has provided the answer actually is the topic expert to the relevant answer or not. Usually, an answer is upvoted but technically and logically it is not correct. There is an increasing failure rate because it becomes difficult to rank the expert of question. Because these sites ranked the user on the basis of the answer he has posted and don't know whether the answer is wrong or not.

3.2 Proposed System

To resolve the problems proposed above, our idea is to develop a platform where a user can find a valid and authentic answer according to a relevant topic and find the expert of that answer. The expert rank method assigns a score to probable answerers based on their likelihood of providing satisfactory responses. The authors will be identified according to their degree of relevance to the question.

The quality of the answer will be identified by using text analysis. We will use the Latent Dirichlet Allocation model to compute the content relevance part and the PageRank algorithm for the expert network part. In content analysis, the similarity between question and answer evaluates the expertise of the user and in social network analysis (SNA), we will use the PageRank algorithm. Link analysis algorithm is considered significant and is adopted in the research of expert ranking. The PageRank-based expert ranking algorithm outperforms other algorithms in social QA forums. PageRank is used primarily for ranking web pages in online search results. Latent Dirichlet Allocation model will also be employed to calculate the similarity coefficient of the candidate profile.
This will help to improve the reputation of the user on the basis of valid answers, content quality of any particular answer, and user participation in already developed blogs, Community Question Answering forum or any Informative Facebook group, etc.

3.3 Requirement Specifications

Requirement Specifications is a detailed description of the software that tells what the software will do and how it will perform in the future. It also describes the needs of stakeholders (persons involve in software e.g., users, client, developers, etc.) and functionalities that are capable of fulfilling the requirement of stakeholders. It defines the purpose of our product and also describes what we are going to make. A good requirement specification describes how software will interact with the person and how a computer system reacts with software. It ensures that during the development process developing team has to complete the requirements of the client in a good manner and manage all the requirements according to the life cycle of the project. The requirement specification of our project is described below in detail.

3.3.1 Functional Requirements

The functional requirements are concerned with the functional behavior of the system. Functional requirements specify a function or set of functions that a system or system component must be able to perform. The functionality to be made available to the users of the system is discussed below.

1. **Requirement 1:** Dataset Scrapping
   Description: The dataset we are going to use is the question-answering dataset. This dataset contains questions asked by several users and answers to those questions given by several authors. The dataset is scrapped from the application programming interface (API) provided by StackOverflow (a popular question-answering forum). The dataset consists of fifty thousand rows where users have posted questions and

get responses from different authors.
Output: Dataset consists of multiple rows.

2. **Requirement 2:** Dataset Pre-processing
   Description: Once the dataset has been scrapped successfully, it would be in uncleaned form. The dataset needs to be pre-processed so that it becomes clean and efficient. There are several natural language processing (NLP) techniques and other data cleaning techniques that are applied to the dataset such as lower casing, removing punctuation, removing HTML tags, etc. When all these operations are performed, the dataset becomes clean and efficient.
   Output: Dataset in pre-processed form.

3. **Requirement 3:** Post Questions/ Get Answer
   Description: The question must be posted by the user or should be available in the form of a dataset so that the user can get multiple answers to that question. Random users post questions of their interest or get answers from different authors according to their knowledge of the topic.
   Output: Multiple answers along with a description.

4. **Requirement 4:** Evaluate Content of Questions
   Description: The user can see the content of a particular question posted by different users. The content should be relevant to the problem asked. Different content relevance techniques will be applied to validate the topic.
   Output: Topic Model.

5. **Requirement 5:**Finding Appropriate Author
   Description: The user should get a valid and accurate answer from the author who is an expert of the relevant topic. The appropriate answer is evaluated by applying link analysis and content analysis. The link analysis analyzes the threads of each author and content analysis checks the quality of the answer. Hence the answer will be ranked according to these techniques and the author will be assigned a badge that ranks the author of the answer.
   Output: Valid/Authentic Answer.

3.3.2 Non-functional Requirements

Non-functional requirements are the overall required attributes of the system that are mainly important for every project which includes performance, availability, portability, efficiency, or understandability of any project. In this project, we have these non-functional requirements.

1. Talking about the Availability of our project, any user can post questions and any author can write answers on any topic according to his knowledge.
2. Project does not require any specific hardware or software. It can run on any system with proper Question answering forums or blogs.
3. It can be modifiable if any person wants to change the ranking criteria or any other functionality according to his choice.
4. It does not require any specific tools for running. It is easily operatable and controllable.
5. It is Scalable because it is easily adapted by any system and a person can add or remove things by his choice.
6. Talking about the usability of the project or system. It can be used by multiple users to achieve specific goals with efficiency.
7. It is easily understandable by any user who knows about any social Blogs or any Question answering forum.
8. Users can provide feedback if use the services of our product and want to change something.

3.3.3 Domain Requirements

Domain requirements specify the characteristics of a particular category or domain of the project. The basic function that is to be performed under that category is pointed out. These requirements are not user-specific because they are identified from the domain model of the project. The domain requirements of our project are discussed below.

1. The dataset that is to be used for implementing the system should be cleaned such that it should not contain any noise.
2. When natural language processing techniques and algorithms are applied to the dataset, the dataset will be able enough so that link and content analysis techniques can be applied.

3. To validate answers, the content of the answer provided by the author should be relevant to the problem asked. So, content relevance ensures that answers should not be vague.

4. The author's reputation also matters in order to ensure answer validity. Hence link analysis makes sure that whether the author has enough knowledge of the problem or not.

5. The resulting system can be integrated with many discussion forums such as Facebook groups etc.

3.4 Use Case Diagram/ Descriptive Use Case

The following use case diagram defines how dataset has been scrapped. This dataset belongs to stackoverflow application programming interface API. After querying on the database schema, the dataset is scrapped.

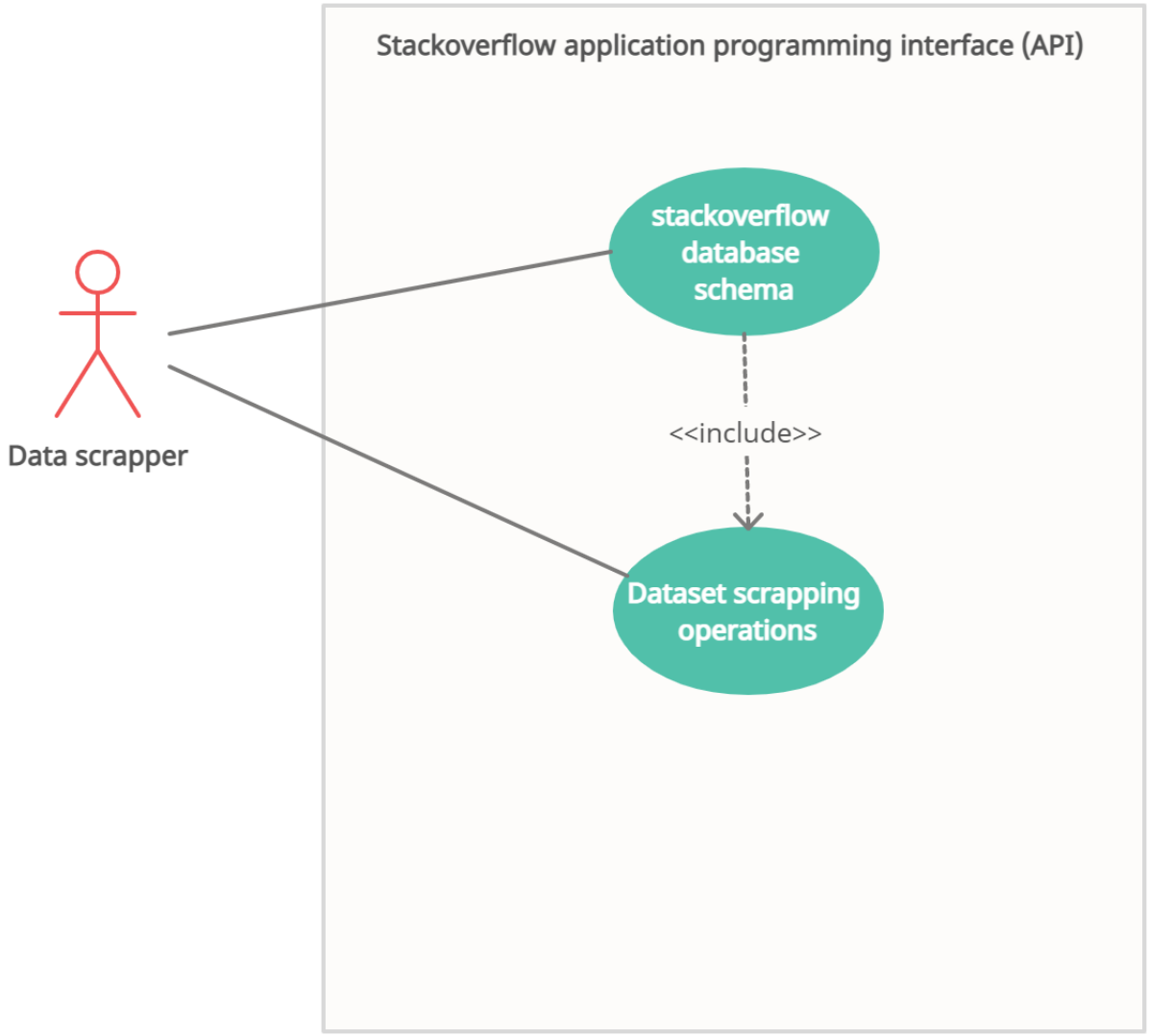

Figure 3.1: Use-case Dataset Scrapping

Table 3.1: System Scrapping Use-Case

Name of use-case	**Scrapping The Dataset**
Use case ID	SCR-BU01
Actors:	Data Scrapper
Description:	This use case describes how the data scrapper has scrapped the dataset. The scrapper interacts with the database schema provided by StackOverflow. On that schema, the query is applied and the dataset consisting of fifty thousand rows has been extracted.
Basic flow:	• The scrapper looks for the API. • The scrapper writes code to retrieve desired dataset. • The dataset is successfully scrapped.
Alternative flow:	**Database schema:** • The database schema contains questions and their answers. • The data scrapper applies operation on the schema.
Pre-condition:	The dataset is not scrapped.
Course Event:	**Actor Actions** 1. The data scrapper successfully scrapped the dataset.
	System Response 2. The data scrapper writes code for retrieving the dataset. 3. The actor runs the database query. 4. The query produced the desired dataset. 5. The dataset is extracted and downloaded in CSV format.
Post-condition:	The dataset extracted consists of fifty thousand rows which contains questions along with their answers etc.

3.5 System Use Case

The following use case diagram is the complete system diagram. It contains all the activities or functional requirements that are to be completed. There are two actors called user and author who coordinate with the system. Both actors are dependent on each other. There are different use cases that belong to each actor. The complete description of this use case is described below.

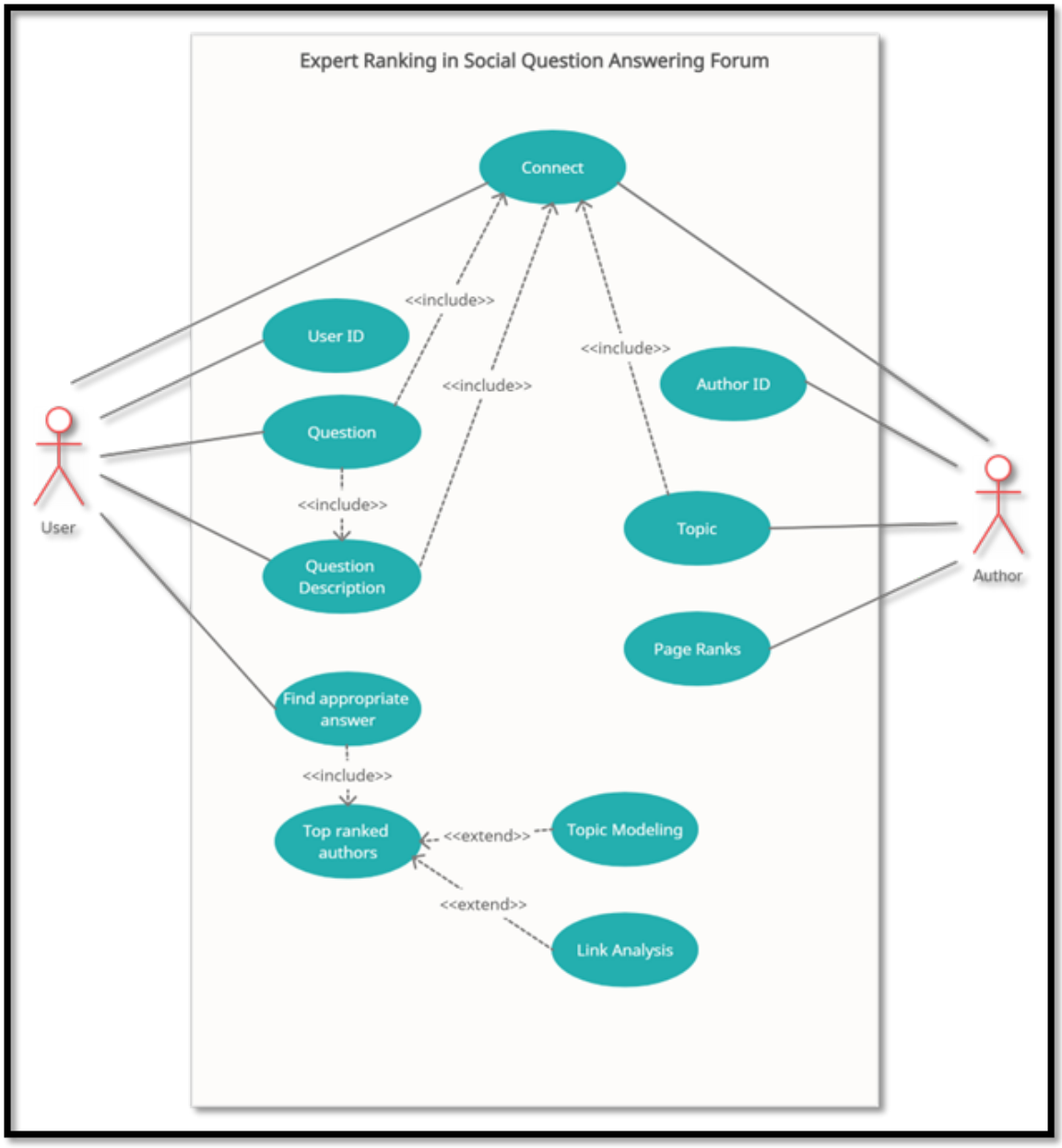

Figure 3.2: Use-case Diagram

Table 3.2: System Use-Case

Name of use-case	**Expert Ranking in Social Question Answer Forum**
Use case ID	ER-BU01
Actors:	User, Authors.
Description:	Use case describes the interaction of different user with any social question answering forum. Random user post question of their interest and can also get answers according to the knowledge of the relevant topic and can see Total Number of Answer from different authors. Users can see the content of particular answer provided by different authors. Users get a valid and accurate answer from the author who is expert of the relevant topic.
Basic flow:	• The user post question and its details. • The user sees the total number of answers. • The user sees the content of answer.
Alternative flow:	**Author:** • The Author see the question posted by users. • The author gives answer of their interest.
	Forum Management: • Forum management provides the question. • It provides different answers of different users with expert priority. • It also shows the total number of answers posted by different authors. • It also provides the content of the answer.
Pre-condition:	No users are registered to our application.
Course Event:	**Actor Actions** 1. The user post question.
	System Response 2. The question has been posted. 3. The forum management match the answer with question subject. 4. It than add answer to particular question. 5. Answer is ranked on the basis of author badge priority, link analysis, content analysis and some other algorithm. 6. Expert Answer content is than shoe to user after all steps.
Post-condition:	The User gets the valid and accurate answer of the expert author.

3.5.1 Post Question/ Add Description:

This use case describes that a user (actor) can post a single question at a time and the user who posted the question should provide the description of the problem asked. The problem then will be answered by different authors.

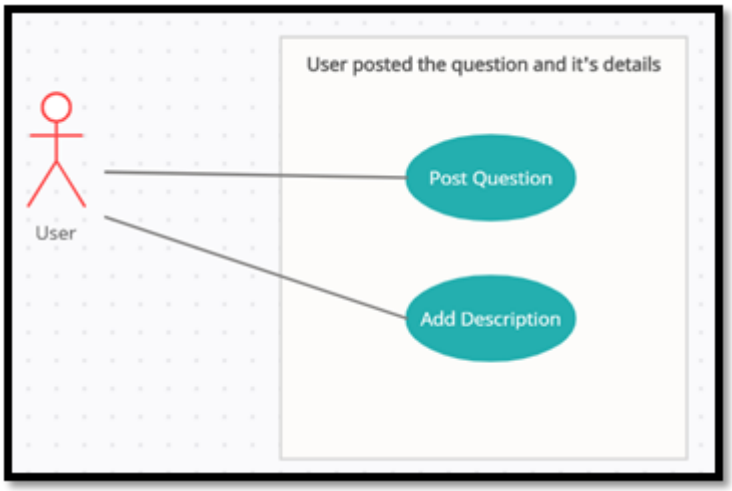

Figure 3.3: Use-case PostQuestion

Table 3.3: Use-Case PostQuestion

Name of use-case	**Post Question/ Add Description**
Use case ID	ER-BU02
Actors:	User
Description:	The use case describes the interaction of the different users with any social question answering forum. Random users can post the question of his/her interest and can get answers provided by answerers according to the knowledge of the relevant topic. The user adds the title and body (detail) of the question to be posted.
Basic flow:	• The user posts the question. • The user adds a description to any Question.
Alternative flow:	**Author:** • The Author see the question posted by users. • The author gives answer of their interest.
	User: • User post Question according to his choice. • Question contain description.
Pre-condition:	No users are registered to our application.
Course Event:	**Actor Actions** 1. The user has posted the question
	System Response 2. The user adds the description of the question. 3. The forum management added a question.
Post-condition:	The User successfully posted the question.

3.5.2 Responses

The user and author relationship are dependent on each other. The problem asked by the user is answered by the author. Multiple authors can provide answers to the problem asked. Then the best suitable answer can be ranked.

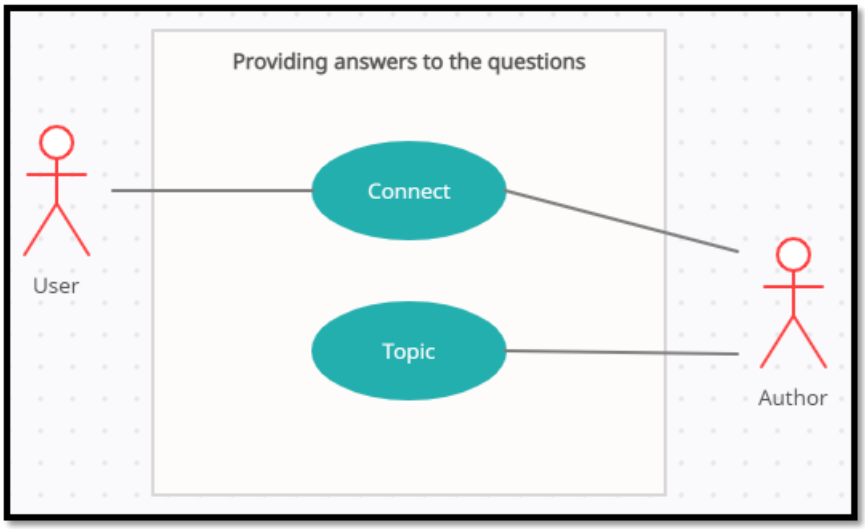

Figure 3.4: Use-case Responses

Table 3.4: Use-Case Manage Responses

Name of use-case	**Responses**
Use case ID	ER-BU03
Actors:	User, Author.
Description:	The authors can look for the posted questions and can give their feedback/answer to the question according to the description of the question.
Basic flow:	• Search questions to be answered. • The answer's content should also be similar/relevant to the question asked.
Alternative flow:	**Author:** • The Author sees the question posted by users. • The author gives answer of their interest.
Pre-condition:	The QA Forum is running.
Course Event:	**Actor Actions** 1. The Question is posted by user is received by System.
	System Response 2. The question is now readable to the author. 3. Author provide answer to question. 4. The forum management match the answer with question topic. 5. It than adds answer to particular question.
Post-condition:	The Question and Answer are posted by respective actors.

3.5.3 Find Appropriate Answer:

Since multiple authors have provided different answers. The link and content analysis techniques validate the quality of the answer. When these techniques are applied, the author is assigned a badge that specifies the best quality answer to the problem asked. The user can rank whose answer is best.

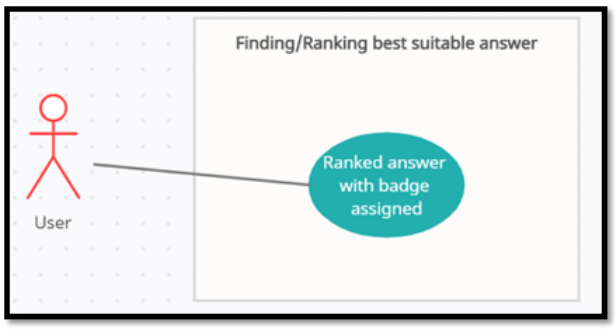

Figure 3.5: Use-case Ranking

Table 3.5: Use-Case Appropriate Answers

Name of use-case	**Find Appropriate Answer**
Use case ID	ER-BU05
Actors:	User, Author.
Description:	Users get a valid and accurate answer from the answerer who is expert to the relevant topic of question.
Basic flow:	• The user gets valid and accurate Answer. • The user sees the expert rank.
Alternative flow:	• The user get answer on basis of expert badge priority. • After Link analysis between Question-and-Answer best answer will provided.
Pre-condition:	No users are registered to our application.
Course Event:	**Actor Actions** 1. The user finds Appropriate Answer.
	System Response 2. The best Answer is provided with higher ranked. 3. The forum management match the answer with question's description using link analysis and content analysis. 4. : Answer is ranked on the basis of author badge priority. 5. Expert Answer content is than show to user after all steps.
Post-condition:	The User gets the valid and accurate answer of the expert author.

Chapter 4

Design

The section describes the architecture of our proposed system. Data preprocessing steps include the cleaning of dataset. It includes different techniques like stop words removing, stemming of the words are performed (rewrite insignificant words and rewrite them in their root form), removing html tags etc. Different Natural Language Processing (NLP) techniques are applied on a dataset which we have extract. Text and link analysis will be done using natural language processing techniques such as topic modeling, named entity recognition which will identify on which specific topic user is talking about. Talking about Model Training, we will use the attributes of this dataset to train our machine learning algorithm which will help us to identify and learn good values for all the attributes.
The system is going to find an expert by the help of expert rank algorithm which is based on two different methods Expert relevance and Expert authority. First method expert relevance is document-based method in which content-based relevance is going to use. By this method we rank the author/expert on the basis of his/her knowledge on a particular topic in Question Answering forum. Second method is based on both document-based and Social-network based. It is the hybrid method of both of these document-based and Social-network (expert authority) based in which we use PageRank algorithm to rank the authors/expert. PageRank algorithm will help to perform link analysis of the expert based on expert authority and expert relevance. Top answerer will then be identified by combining text and link analysis.

4.1 System Architecture

System architecture describes the basis for the proposed system. In our system, the architecture defines what basic building blocks are required to develop the system. The proposed system contains the question-answering dataset. This dataset provides the basis

for applying all the operations and techniques to generate the desired outcome. This dataset should be preprocessed to obtain the desired output. The preprocessing involves data cleaning and some natural language processing techniques. When the data is preprocessed, we can apply the techniques mentioned in the diagram such as text analysis and link analysis. This modeling of the dataset will result in the User's ranking for the particular answer.

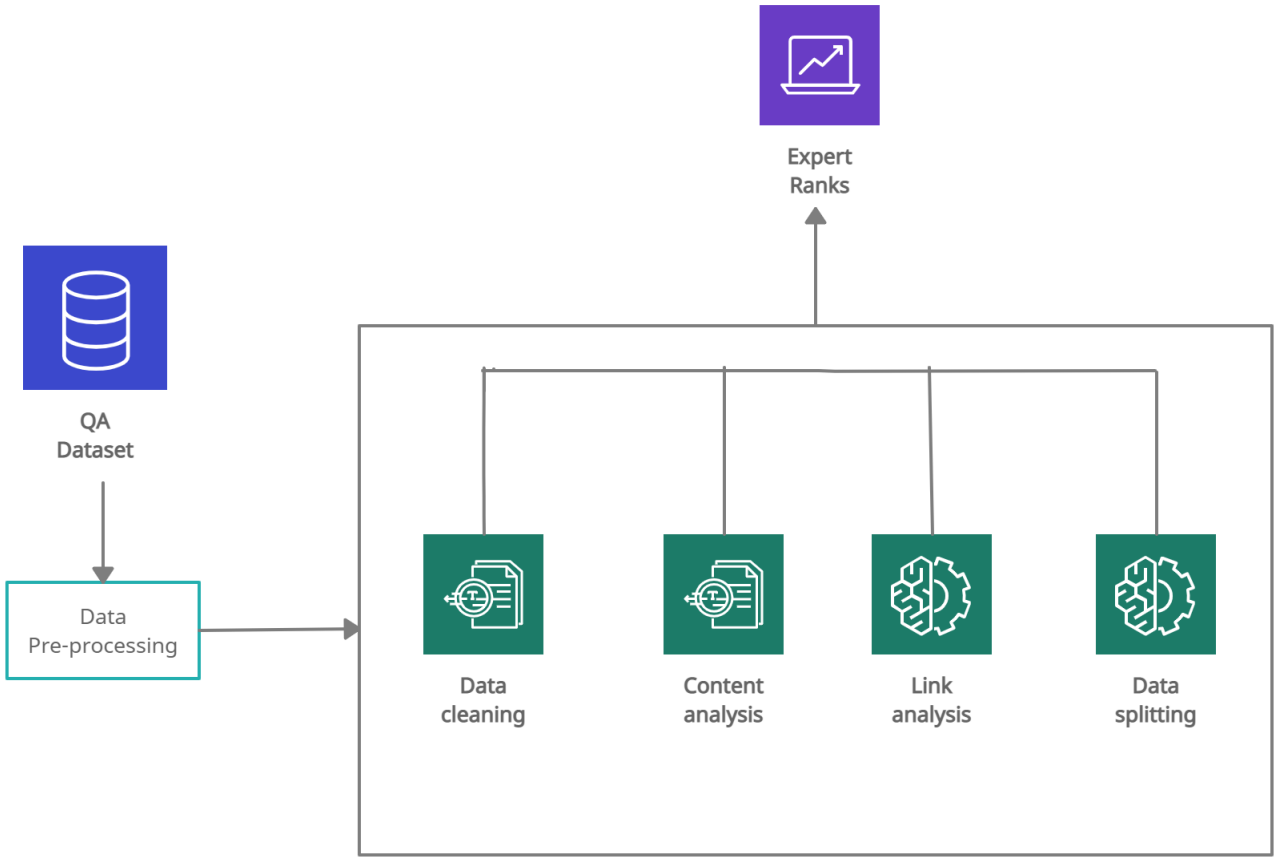

Figure 4.1: Architecture Diagram

The dataset used for the project is scrapped from the application programming interface (API) provided by stackoverflow. The dataset is extracted by writing the code for it and then the desired dataset of fifty thousand rows is retrieved which contains questions and the persons who have posted the questions and the answers to those questions along with the authors of the answers.

The pre-processing of the dataset involves natural language processing techniques and some python algorithms to make data in the cleaned form. Natural language processing applies to the textual form of the data such as data lower casing, removal of punctuation, removal of HTML tags, etc. After applying all these techniques, the dataset is able to get

operations on it.

The content analysis techniques such as latent dirichlet allocation is applied on the dataset to analyze and identify topics from the questions and answers. it will help the system to see whether the answers are relevant to the questions asked or not.

The link analysis algorithm such as PageRank algorithm is used to find links of each other at different questions and with other authors. This will help the system to identify whether the author is expert of the question or not. After analyzing all these techniques, the system will find out the expert to the particular question.

4.2 Design Constraints

Talking about the design constraint of the project, our expert finding algorithm is based on the hybrid Expert finding technique. There are many Social Question Answering forum that ranks user on the basis of self-classification and Document-based relevance only. Also, there are also many Communities Question Answering forum that ranks authors on the basis of self-disclosed information and based on social network analysis so they are not able to find an expert on a specific topic.
Our proposed system used the hybrid expert finding technique because there are many recent systems that start considering both social network analysis (PageRank algorithm) and Document-based relevance (involve text analysis technique to capture authors expertise). In this way, it is easy for any Community Question Answering forum to find the expert who has the best knowledge of the particular topic and is also socially active in a different network of communities. The technique can be applied to different Social Question Answering forums or different knowledge repositories having low Information quality but rich in social media. It Dynamically ranks experts in an area specified by any search query.

4.2.1 Programming Language/Techniques

The Programming langue that we are use in our project is Python Programming language. Python is used because by using it we can easily manage our big dataset in preprocessing step, text analysis or any other data related work. Python provides many libraries for performing different tasks such as NLTK (natural language toolkit) for applying natural language processing techniques.

- TextBlob
- Pandas
- Numpy

- Matplotlib
- Pytextrank
- Scikit-learn
- Spacy
- Sklearn

These are some of the libraries which will use in our system. Their are more techniques that will be applied to the system. So the libraries according to those techniques will be used then.

4.2.2 Resources Used

The dataset used for the system is the basic resource for the project. A Computer system having some tools like visual studio code, Jupyter notebook, microsoft excel etc. and capable of running these tools smoothly is enough for this project. The system also requires to have solid state drive which is good for performance measures.

4.2.3 Assumptions and Dependencies

It is an independent software and that has been used by any Social Question Answer forum or Blogs to find the expert author. We assume that this algorithm in future will be used in different Mobile applications blogs or different informative sites that are running on phones.

4.3 Design Methodology

The dataset that is for our system is related to community forums such as stack overflow. The preprocessing steps include data cleaning, removing punctuation marks from the strings, removing stop words, stemming of words, removing different HTML tags, etc. The two main approaches are used to find the expert. These approaches are social network analysis and Document-based Relevance. We used text analysis and Latent Dirichlet Allocation model in document-based relevance to capture author expertise and PageRank Algorithm in expert social network analysis. In content analysis, the similarity between question and answer evaluates the expertise of the user and in Link analysis algorithm is considered significant and is adopted in the research of expert ranking. The PageRank-based expert ranking algorithm outperforms other algorithms in social QA forums. PageRank is largely used to rank web sites in search engine results. The LDA model will also be used to calculate the candidate profile's similarity coefficient. Finally, the two portions are linked together using a cascade strategy.

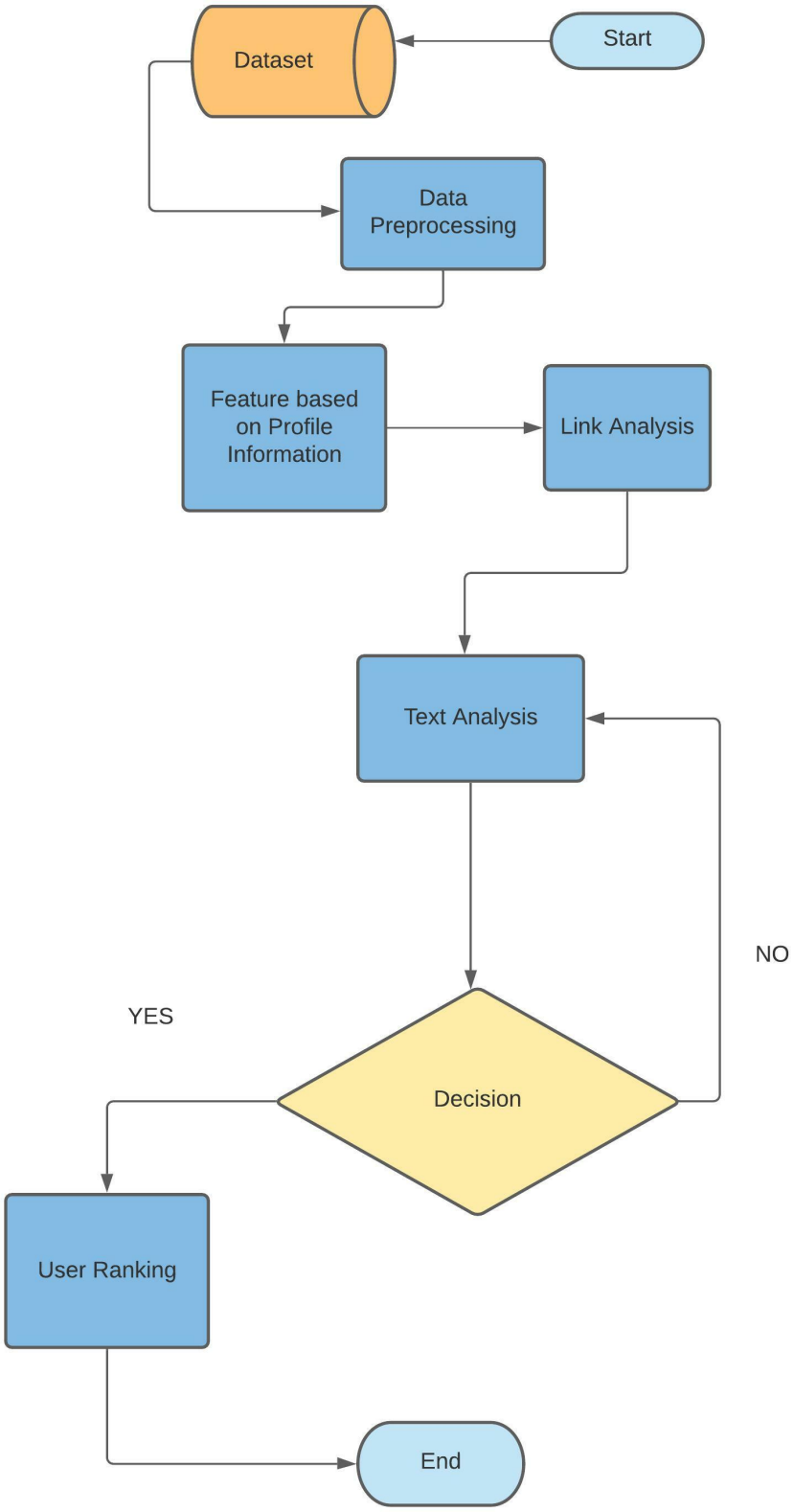

Figure 4.2: Flowchart Diagram

4.4 High Level Design

This section expands on the elements described in the Architecture section of any system. High-level design is the way of modeling the group of system elements in a different number of views. The high-level design of our proposed system includes diagrams that depict our system. The diagrams along with their description are discussed below:

4.4.1 Sequence Diagram

The sequence diagram is the way of describing the different activities of a system in steps and also describes the behavior of a system with users including the result of every single activity that has been performed by a user. In this sequence diagram of our proposed system, we assume that when users attract to any Community question-answer forum. When a user posts a random question, the question with description will be received by an author and he will post his answer according to his/her choice. These all steps are done in multiple sequences/levels. And users get a different answer from different users who are most authentic. This authentic and expert answer is provided to the user after the authors' ranking. Each step below in the sequence diagram describe the overall working of our proposed system.

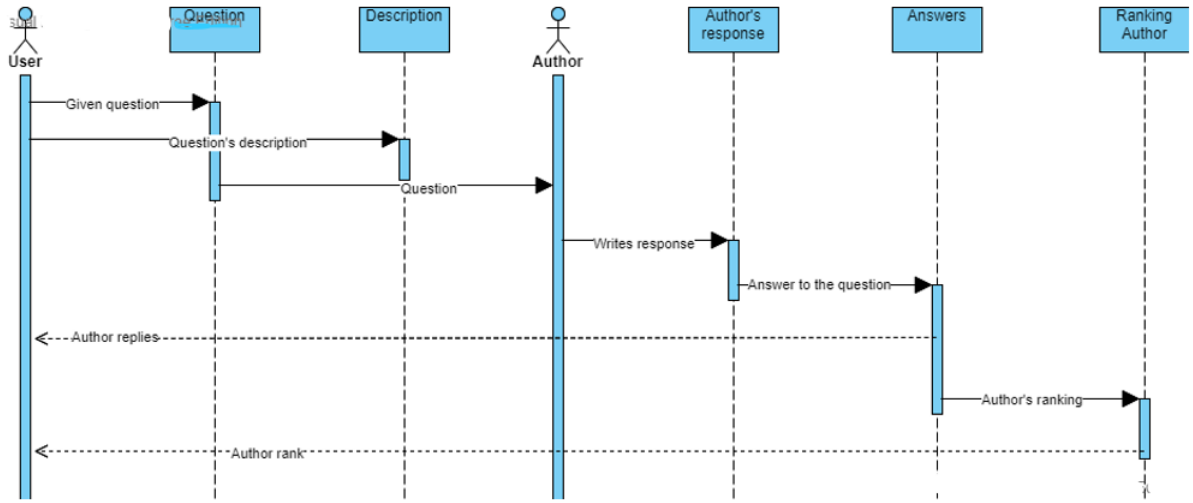

Figure 4.3: Sequence Diagram

4.4.1.1 Dataset Scrapping Sequence Diagram

This sequence diagram describes the sequence of scrapping the dataset. The dataset for the project is scrapped through the application programming interface (API) provided by stackoverflow. The scrapper writes the query to get the desired dataset. The dataset of fifty thousand rows is retrieved by the query and the dataset is made available to the scrapper.

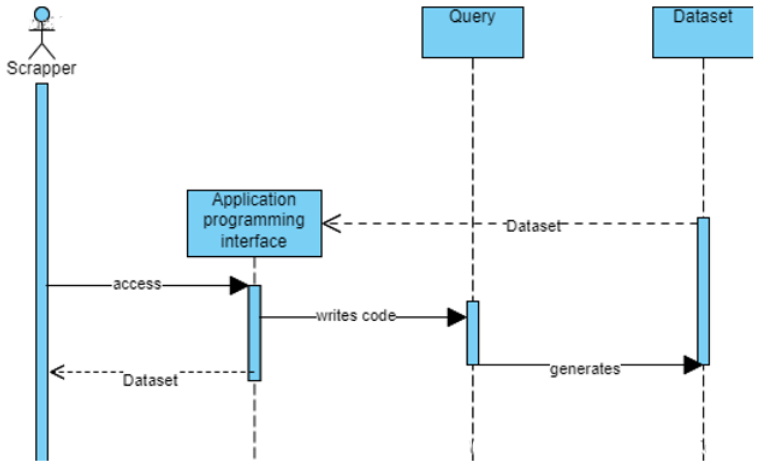

Figure 4.4: Scrapping Sequence Diagram

4.4.1.2 Questions Sequence Diagram

This section describes the flow of question posted by the user. The user asks the problem and adds description of the problem. This question can now be answered by number of authors. The user can get responses against the question.

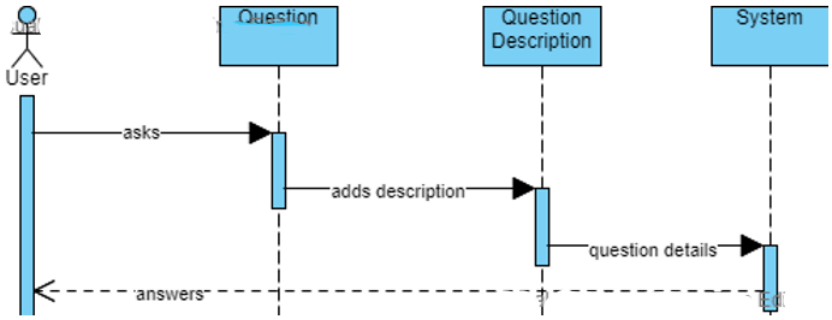

Figure 4.5: Question Sequence Diagram

4.4.1.3 Answers Sequence Diagram

This sequence diagram describes how the author respond to the user's questions. The authors get the questions from the system. Since many authors can respond to the single question. So the question is answered by number of authors and user receive multiple answers.

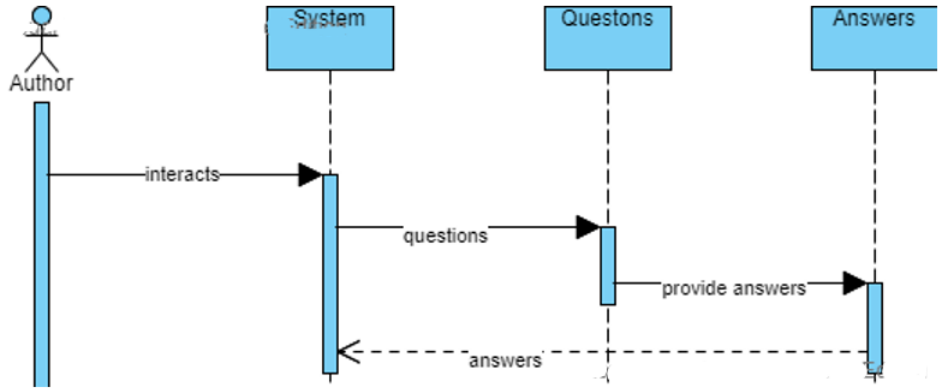

Figure 4.6: Answers Sequence Diagram

4.4.1.4 Expert Ranks Sequence Diagram

The user interacts with the system. The user can see how many answers are provided against the question asked. The user can also evaluate who is the expert of among the answers. The one's answer relevant to the problem asked and whose links are strong will be the expert of that question.

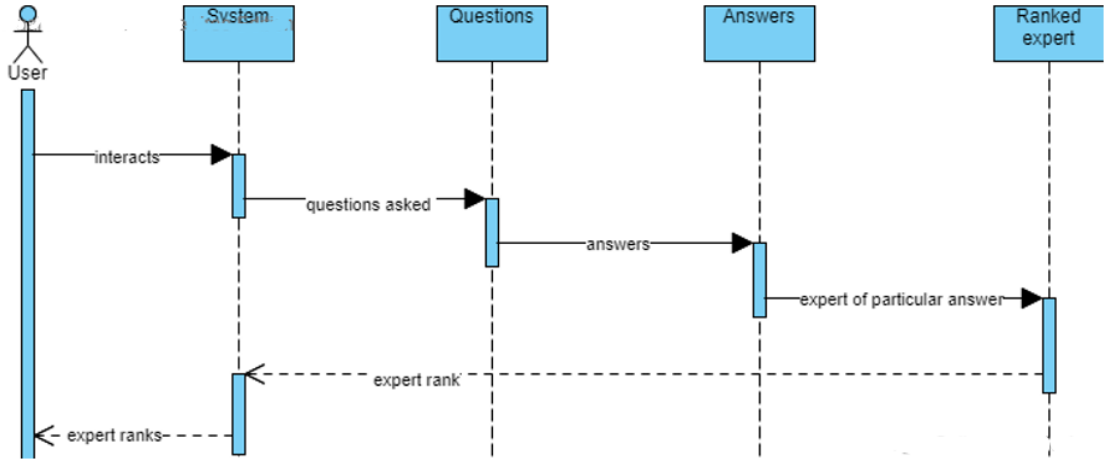

Figure 4.7: Ranks Sequence Diagram

4.4.2 Context Diagram

Context diagrams depict how a system interacts with other actors (external factors) with whom it is supposed to engage. System context diagrams can help understand the context of which the system will be part. A context diagram shows the entire system as a single process. In this Context diagram of our proposed system, the overall working of a system is shown in a single unit. System/process are shown in a square round like user and authors who are responsible for box posting question and answer and also see posted answer and question. And in center circle box show the overall working of the ranking system in online communities. The data flow of a system is shown in arrows.

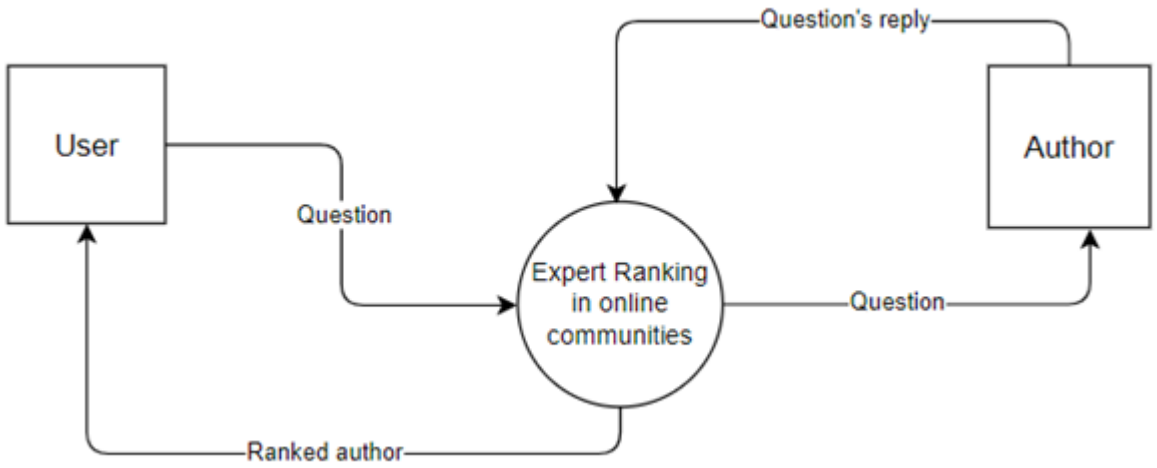

Figure 4.8: Context Diagram

4.4.3 Class Diagram

A class diagram is used to describe classes and their relationships in software design and modelling. The graphic depicts the names of the classes, their attributes, and the relationships between them. In our proposed system we have a class diagram of our system having different attributes and their relations with each other. The scrapping class scraps the data from the API and retrieves the dataset for the system. The ranks class ranks the answer based on content and link analysis that are composed into separate classes. There we have different classes like main class expert rank which is directly linked with User, Author, and Natural language processing. The attribute of the expert rank class is Question, Answer, and Ranks which are directly connected to User having (User Id, Username) and Question class having (Title and Description of the question) and similarly for answer class that is directly connected to authors class and author class. The Natural language processing class includes different operations like stop words removing, removing punctuation marks, removing the casing, removing URLs, removing HTML tags, etc. after that we have tokenizer class which is directly connected to word tokenize and sentence tokenize class having attribute (Question, Question Body and Answer Body). Word tokenizer and Sentence tokenizer have one-to-many relation with Tokenizer which show that the same technique is applied on different question having different content.

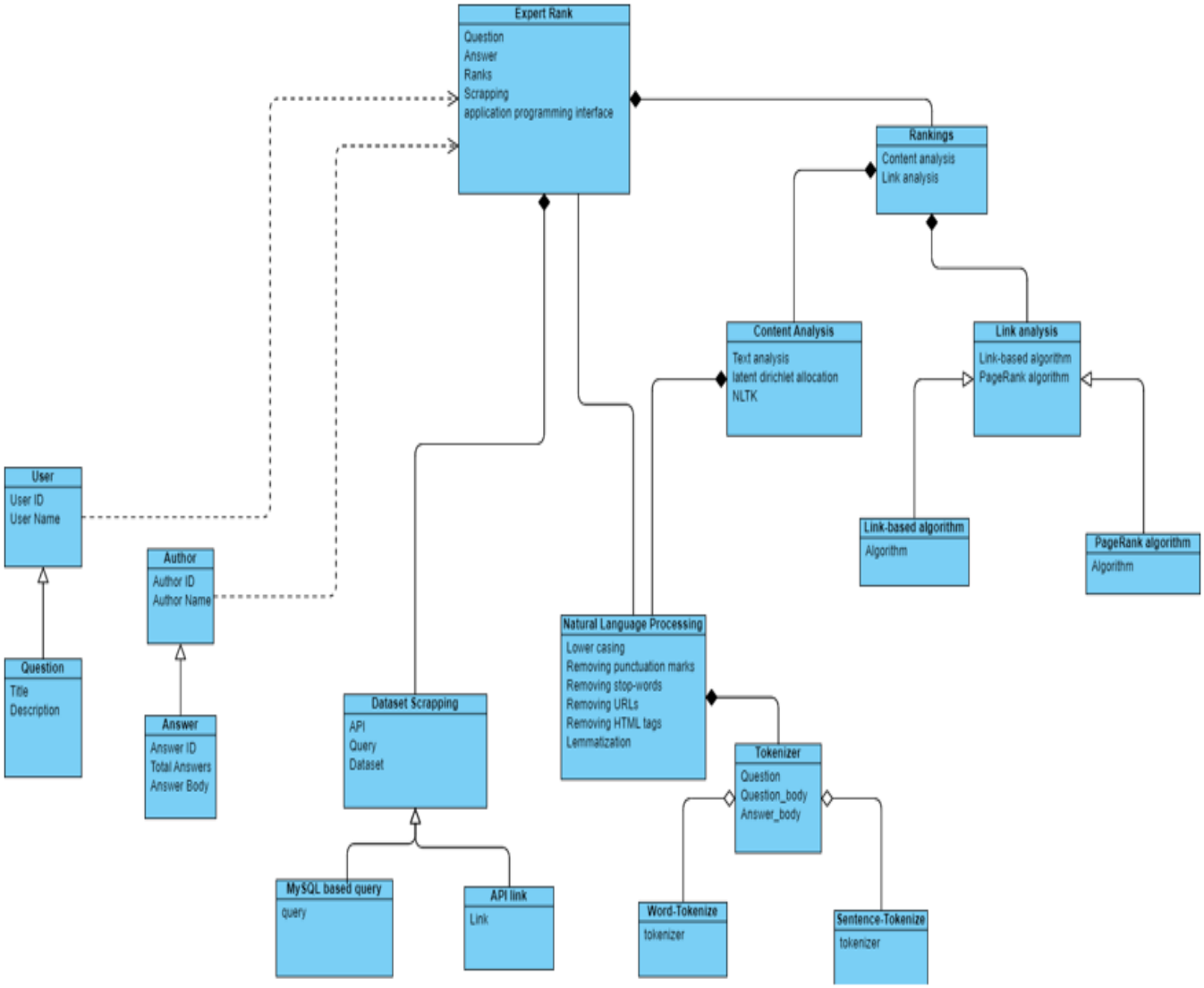

Figure 4.9: Class Diagram

4.5 Low Level Design

The low level design of our proposed system describes the detailed design of the system. In this section, we discussed the fundamental design constraints of our system. We define each module of our system and how that module is dependent on other modules is discussed in this section. The further details are discussed below:

4.5.1 Block Diagram

A block diagram describes the system in which the major parts or functions are represented by blocks connected by lines that shows the relationships of the blocks with each other. The block diagram of our proposed system describes the overall working of a system from start to end. First of all, we have scraped our dataset from stack overflow having different attributes like (post Link, Post ID, User ID, Username, Question Title, Question Body, Authors, Answer). The next block shows the data preprocessing step which shows different operations like stop words removing, removing the casing, removing punctuation marks, removing HTML tags from Question and Answers, etc. Text analysis block which performed different operations like text analysis using different natural language processing techniques like topic modeling, named entity recognition, etc. Then we have a link analysis block in which we use different algorithms like PageRank or expert relevance score etc. to find the relation of authors with answers. And the last block with a Ranking expert shows the top answerer in the graph on the top which is done after link and text analysis. These are the overall functionalities of our proposed system which are shown in form of a block diagram below.

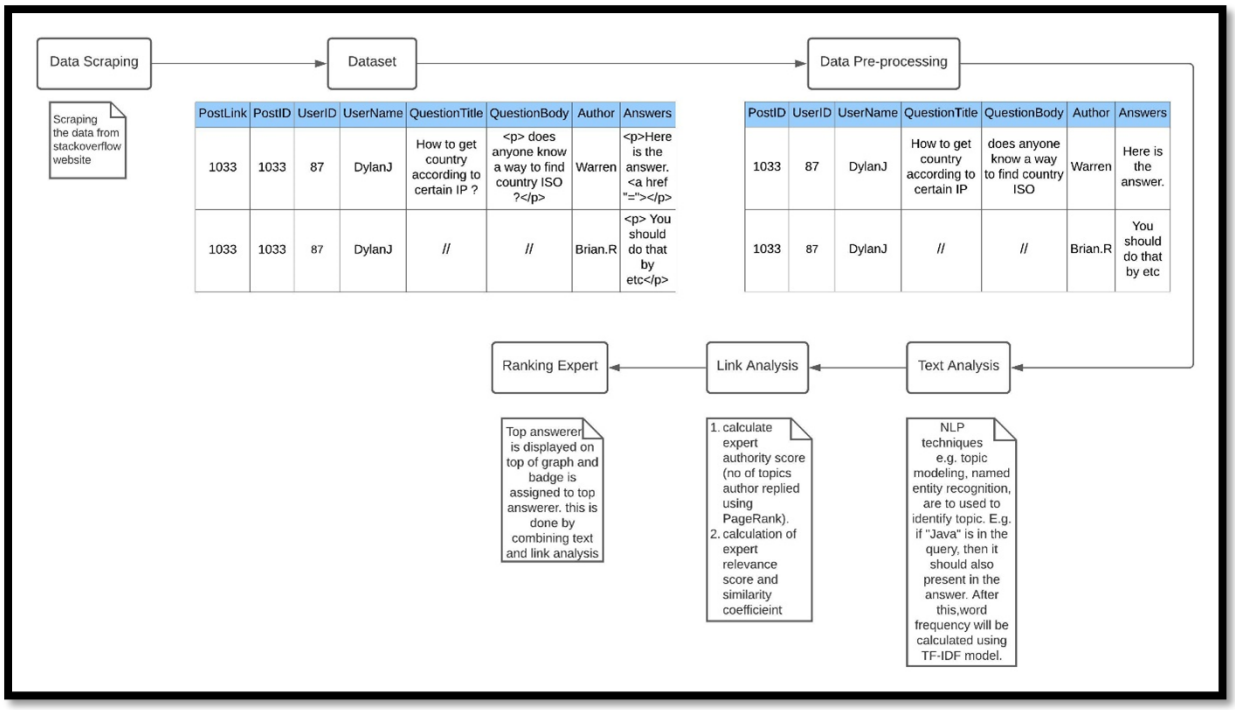

Figure 4.10: Block Diagram

4.6 Graphical User Interface Design

The graphical user interface allows the users to interact with the system through meaningful icons and objects. The graphical user interface of our system is user friendly so that any beginner can understand what is being happening. The GUI is developed on ReactJS which is powerful library of JavaScript for creating responsive web pages. Some of the screen shots of system are mentioned below:

4.6.1 Select Topic

The user selects the topic from the drop-down list. After selecting the topic, the table appears which contains the questions, the questioners, and the authors.

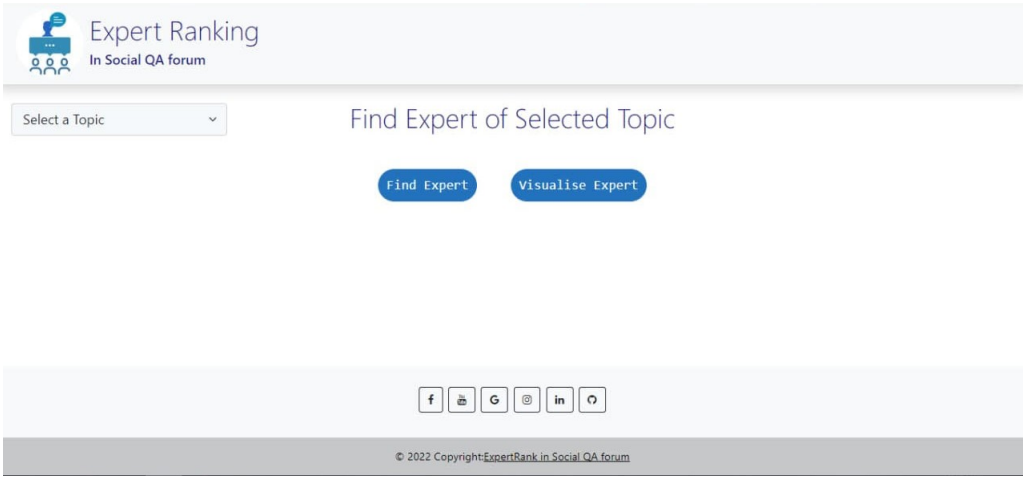

Figure 4.11: Topic Selection

4.6.2 Topic Details

The topic has now been selected from the drop-down list. Below are the details belong to the topic.

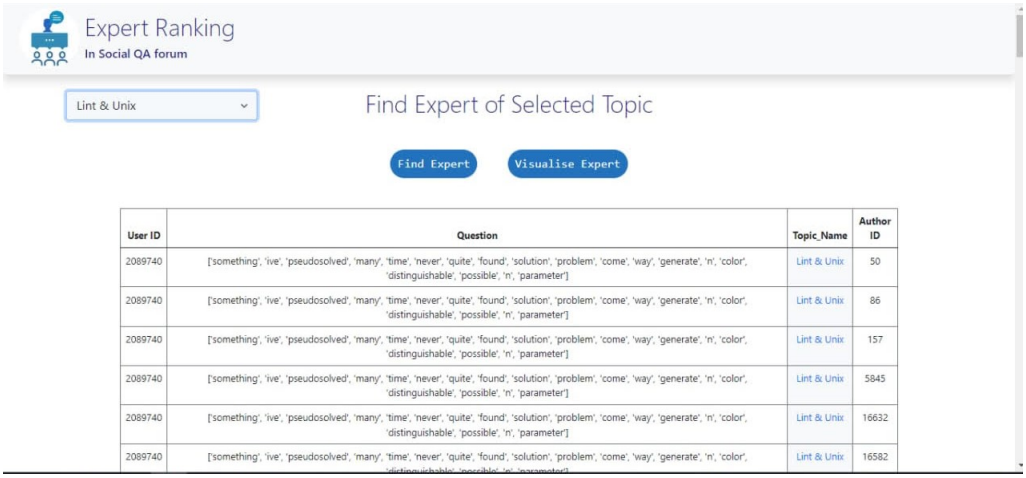

Figure 4.12: Topic Details

4.6.3 Expert Author/s

Their can be more than one experts of the same topic. The PageRank of the authors determines the authenticity of author. The higher the PageRank the value, the more the

chances of being an expert.

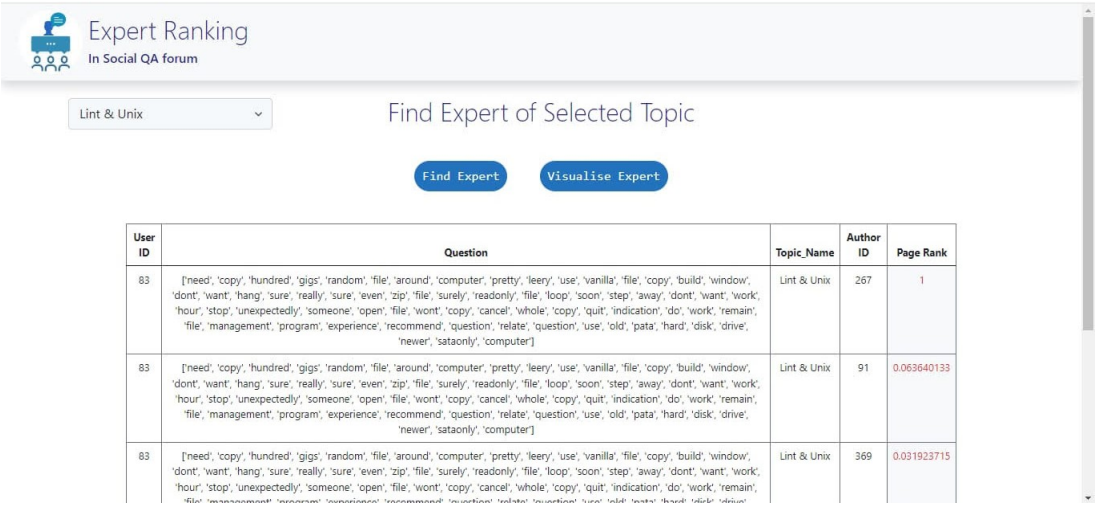

Figure 4.13: Authors

4.6.4 Visualizing Experts

It can be complex to identify who is the expert of certain topic. Here, comes the visualization which helps in finding out who is the expert/s of certain topic. The visualizations are figured below in terms of bar chart and pie chart.

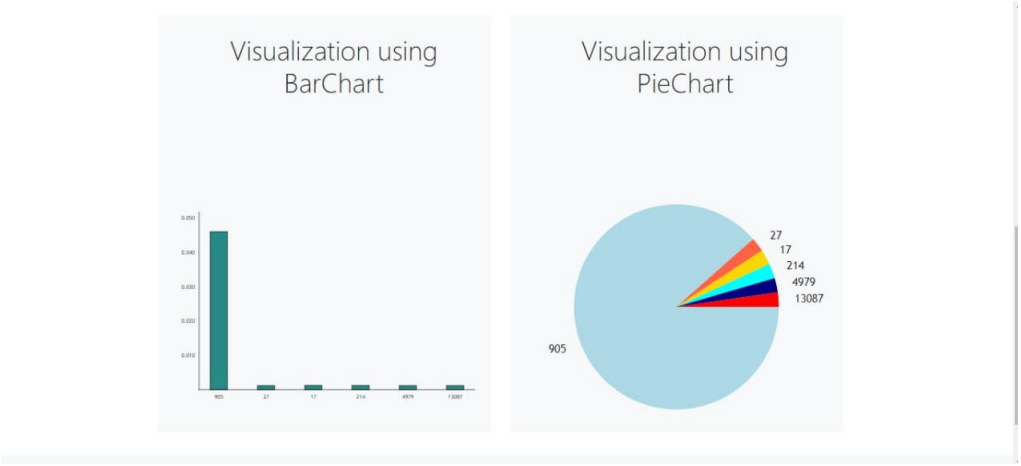

Figure 4.14: Topic Experts

The high bar chart shows the expert authors. Similarly, the pie chart whose area covered is bigger shows the expert author/s.

4.7 External Interfaces

This project is applicable for the question-answering forums. There are a number of online systems such as Quora, Facebook groups etc. where the algorithm build in this project can be applied. This projects targets to those users who posts their problems online and still get confused for the accurate solution. The other systems that already have community forum can integrate this algorithm for better user experience. The other systems should ensures the hardware and software requirement before integrating this system.

Chapter 5

System Implementation

System Implementation is how we define our system should be built according to the given information in the documentation. This process ensures that all the information provided about the project is operated and used according to a quality standard. It describes the overall architecture of the system i.e., How the system should look, what the system should do, and how we built the system. The system implementation of our system involves scrapping the dataset till the final product of what the system produces.

5.1 System Architecture

The system architecture is the model that describes the structure, behavior, and different view of the proposed system. It also describes the internal components of the system and the functionalities of different components of the system. Different tools and technologies are used to implement the proposed system. It also describes different techniques that are used in this project to implement this system. The architecture is based on data fetching then applying some techniques to make the data in a cleaned form. To implement the topic modeling in our system, we applied different algorithms to train our system to make the decision. Social networking algorithms are used to create a social network. Training and testing techniques are used to make a system to take decisions. At the end, the system ranks the potential user.

5.1.1 Data Pre-processing

Data preprocessing is the technique in the data analysis process that converts the original data into such form where the algorithms can be applied. it transforms the raw data into a form that can easily be understandable by any machine or computer system. In this project, we extract a data set from the StackOverflow application programming interface

which includes thousands of data including questions and answers from different users. The dataset contains much raw data including empty rows, many users without their identification number. The preprocessing of the data removes ambiguity and noise and makes it in a form that can be fully understood by the machine.

5.1.2 Applying Natural Language Processing (NLP) Techniques:

The natural language processing technique is the artificial intelligence technique that is used to make human input language understandable to machines. We have applied different natural language processing techniques on our dataset i.e., First of all, we remove all the punctuation marks from the dataset, we also remove all stop words from our dataset. Another natural language processing technique that we applied to our data is lowercasing on textual data which makes all our data into lower case. Then we remove different HTML tags from our dataset that are present with different questions and different answers which do not have any meaning. After that, we also remove different URLs from our dataset occurring in different columns. Then we have applied lemmatization with part of speech on our dataset. It is the process in which inflected forms of a lexeme are grouped into base dictionary forms. Part of speech tagging and lemmatization is the most crucial step of pre-processing. And the final technique that we have used on our dataset is tokenization. It is the process in which we replace our sensitive data from a dataset with some unique identification symbol that retains all information about the data. So, these are all Natural language processing techniques that are used in our system.

5.1.3 Text Modeling

Text modeling is the process to identify a particular topic from the given sentence or paragraphs. Text modeling is used in many community forums as in our system to ensure the system decides on what particular topic the question belongs to. A random user asks the question with a bunch of text inside the question. When applying text modeling, we can easily identify what the user wants to talk about by extracting the topic from the user's question. It is the way to identify patterns of words from different documents in textual material.

To apply topic modeling to the data, the data needs to be split into train and test datasets. We do this split to train the system on a large amount of data so when the system sees any unpredicted input, it can make a decision based on the trained data.

The word ID mappings contain the unique ID for each word. The dictionary is created from the list of sentences. We transform each document to a number form. The Bag of words technique is used which identifies the occurrences of the given word within a document.

Then the coherence score is computed to identify the total number of topics. Topic coherence measures how well a topic is supported by a set of text. The higher the coherence score within a given range of documents, the higher the probability of choosing the number of topics.

Latent Dirichlet Allocation is the most popular unsupervised topic modeling technique. It works with different documents and each document is made up of different words. LDA finds the topic a document belongs to. The probability of each word is calculated and then we sorted the probabilities to represent the topic.

Here are some of the topics highlighted in word clouds. The word cloud is a visual representation of text data.

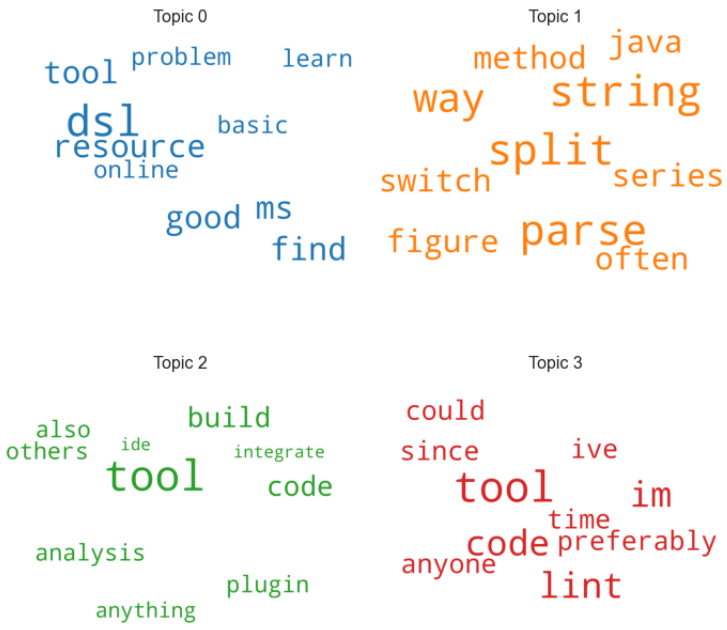

Figure 5.1: Word Cloud Diagram

After identifying different topics for the documents, we assigned the topics by giving the particular term or keyword. Whichever term belongs to the document will be assigned and the topic for each document will also be assigned.

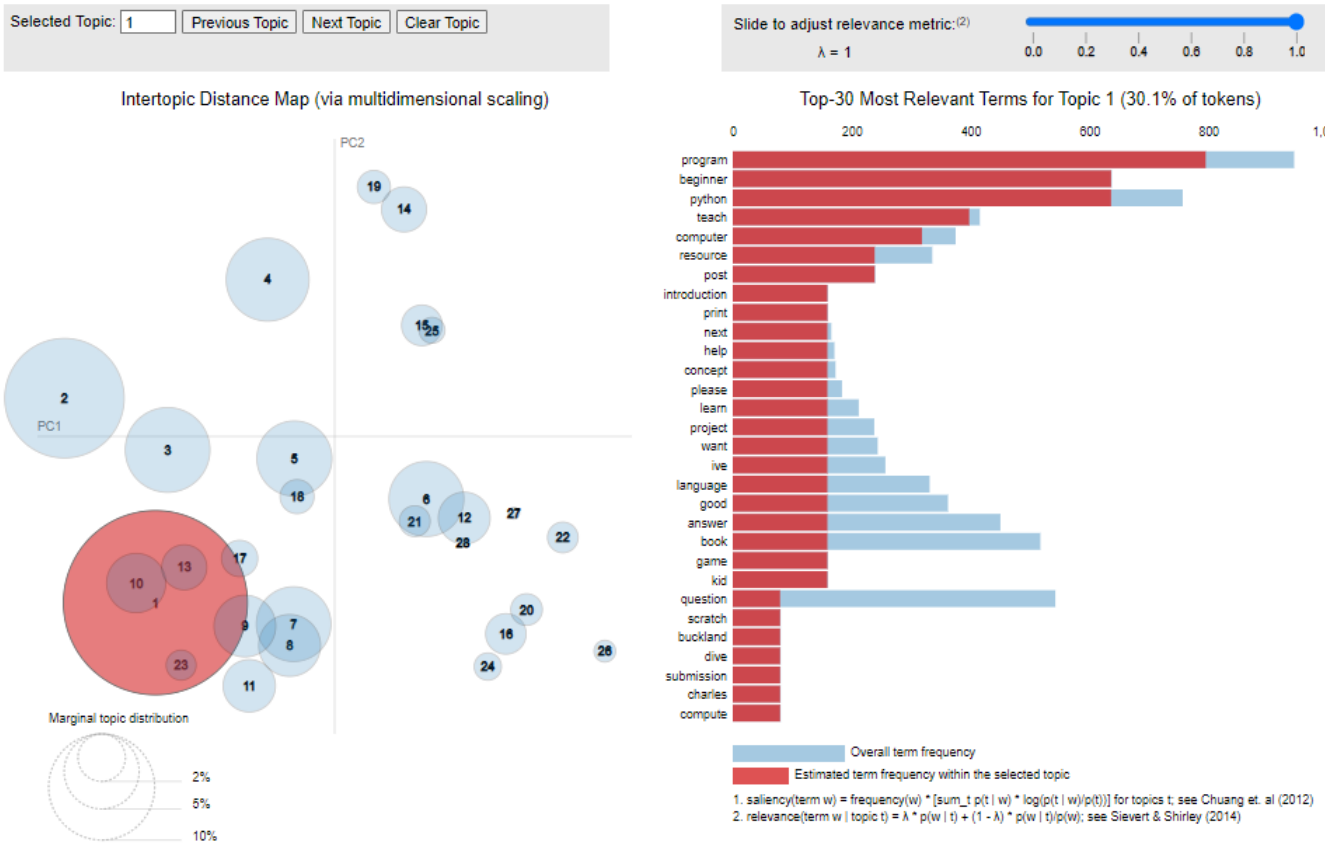

Figure 5.2: Top Most Salient Terms

The topics are assigned by reviewing the top most salient terms produced by Latent Dirichlet Allocation (LDA) model. Each document is reviewed and then the topic name is assigned to that document. Here is the sample of top most salient terms produced by LDA model. The higher the frequency of each word occurring in each document, the higher the chances of picking that word and combining most accurate words to form the relevant topic.

5.1.4 Link Analysis

We have applied Link analysis to our data, Link analysis is the text-analysis technique used to evaluate the relationship between different nodes. In this project, we have applied link analysis by creating a network between the Users (who posted the question) and Authors (who replied to a particular question). After that, we had Identified the connections of each user.

The link analysis algorithm is performed by applying **PageRank** algorithm which is used by google search to rank the web pages. Similar to our project, the PageRank algorithm is used to rank the potential users who have replied to the particular questions and

have string social network. The authority of a network node is determined by the network's connection structure. The goal is for a user's knowledge level to be higher than that of the users he or she can assist. Similar to how the value of a web page is determined by the pages that link to it on Google, the importance of a web page is determined by the pages that link to it.

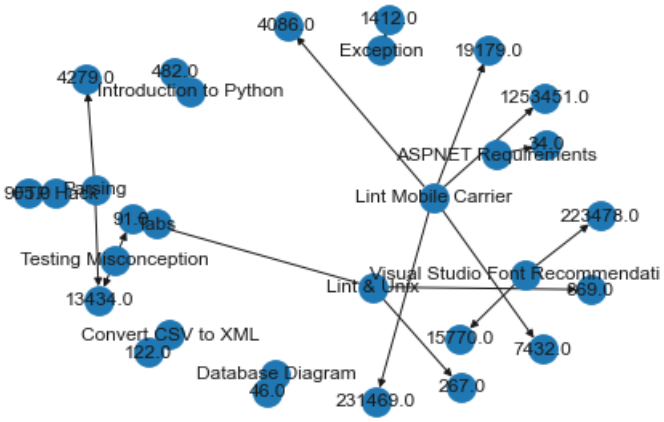

Figure 5.3: Topic Author Relationship Graph

The PageRank algorithm works by creating the graph of the authors who have replied to the particular topic. A topic author relationship is created to make the whole network. In this network, those authors who have replied to the particular topic in many different places have higher PageRank, and those who replied only once have zero PageRank because the answer from that author can be ambiguous or can have weakness in the content. The more the number of connections of the author, the more would be the chances of an expert author.

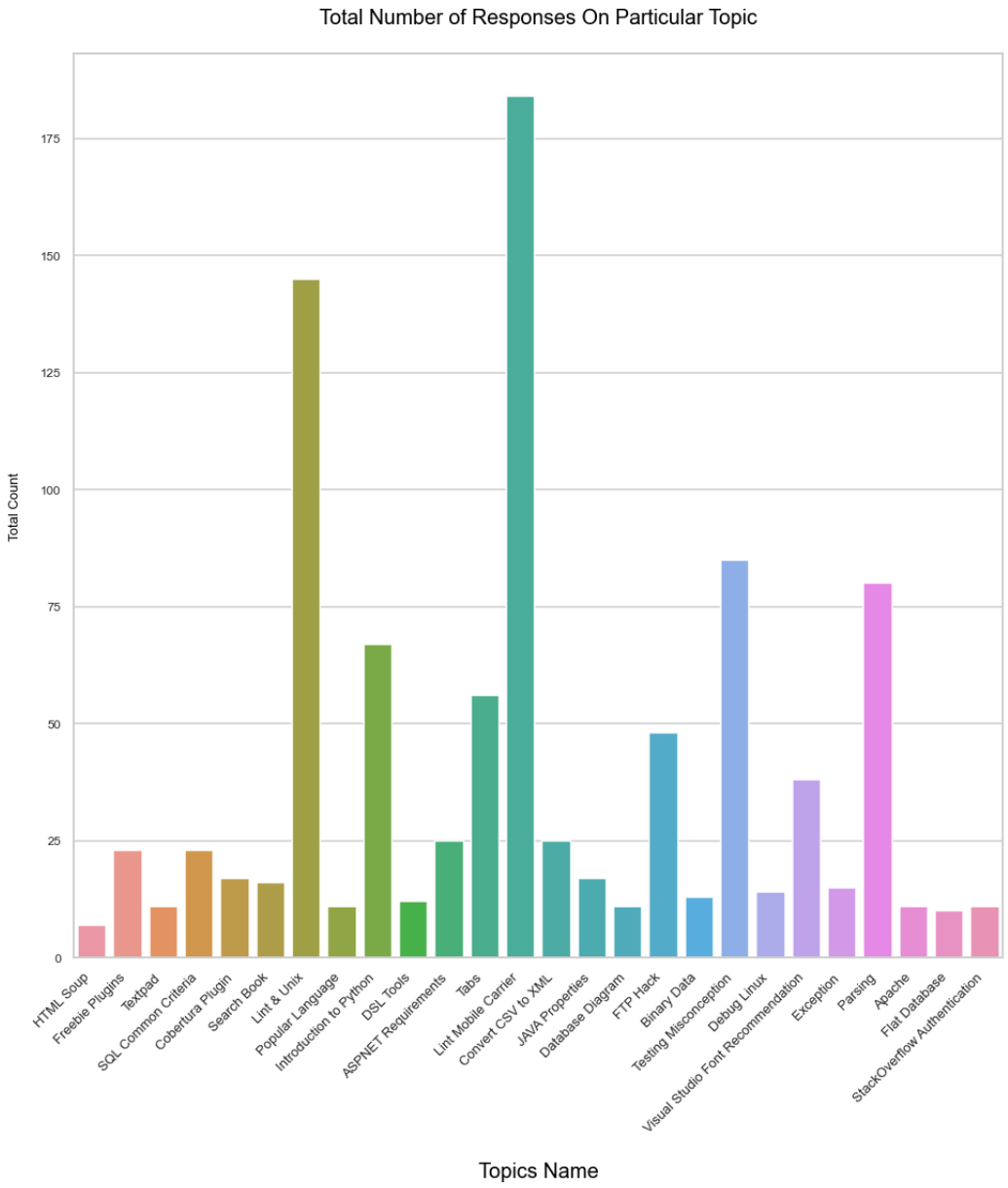

Figure 5.4: Total Responses On Particular Topic

This bar chart shows the number of responses on each topic answered by different authors. The higher the bar chart, the dominant the topic is. The topic whose bar chart is high is showing that this topic has been discussed multiple times in the community.

5.2 Tools and Technologies

Different Tools and technologies are used by different projects. In this project, we also used different tools and technology. The tools and technology contains different libraries and techniques. These libraries involves integrating some algorithms that are to be applied on the dataset.

5.2.1 Libraries

- **Natural Language Toolkit (NLTK)**
  Natural language toolkit is a platform that is used for applying Natural language processing (NLP) techniques on different python programs. These python programs are in human language and different libraries in it are used to change the language to machine understandable language. We also used these libraries in our project when we applied different NLP techniques on dataset.

- **Pandas**
  Pandas is the python library or tool that is used to handle and control relational or labeled data in an easy manner. It supports the different formats of datasets like.csv, JSON, .txt, etc. we also use this library in our project to handle our dataset.

- **NumPy**
  NumPy is the python library that is used to handle different types of arrays and also used to handle many numerical computational problems that were slow. It is also used in our project for different numerical computational problems.

- **Matplotlib**
  It is also a python library is extending from NumPy that is used to plot different graphs on GUI (Graphical User Interface). We also used this library in our project for different functionalities.

- **SpaCy**
  SpaCy is the open-source library that is used in Python programs. This library helps for handling different Natural language processing techniques and we also used this library in our project when we have applied different Natural language processing techniques.

- **Scikit-learn**
  Scikit-learn is a python library that is used for machine learning projects. It also works on predictive modeling and we also used this library on our project.

- **Gensim**
  Genism is a python library that is used in a different machine learning projects. In this

project, we used it for unsupervised topic modeling and different natural language processing technique.

- **NetworkX**
  NetwrokX is a python library that is used in machine learning projects when working with different graphs and networks. We also used NetwrokX library in our project when we are working with different networks.

- **CountVectorizer**
  CountVectorizer is a Skit Learn feature that is used in different machine learning projects for converting text documents into a single string (maybe tokens). We also use this library in our project.

- **Bag-of-word**
  The bag-of-words is a technique used in Natural language processing techniques. By using this we consider or dataset as a bag of words and not consider the grammar. It is also used in different text modeling techniques.

- **WordNetLemmatizer**
  WordNetLemmatizer is a function in python used to do the lemmatization technique. In this technique, we convert a word into base form and we also use this technique in our project.

- **BeautifulSoup**
  BeautifulSoup is a python package is used to fetch data from Html and XML documents. It is mostly used when we are extracting or scraping data from any website. We also used this in our project when we were working on our dataset.

- **nltk.tokenize**
  Nltk.tokenize is the package of python used in different Natural language processing techniques. By using this we divide a different string into substrings. We also used this package in our project.

- **seaborn**
  Seaborn is the python library that is used for data visualization by using matplotlib. Mostly used for making statistical graph. It builds on the top of matplotlib. We also used this in our project for making different graphs.

- **pyLDAvis**
  pyLDAvis is a python library that is used in topic modeling technique for visualizing topic modeling. We used this library in our project when have done topic modeling on our data.

- **CoherenceModel**
  CoherenceModel is used in topic modeling for evaluation of the topic. It is the implementation of the pipeline which includes segmentation, probability estimation, Aggregation, and measuring confirmation. In this project, we used the coherence model when we applied topic modeling.

5.2.2 Frameworks/ Languages

For scrapping the dataset, the structure query language (SQL) is used to extract the dataset. The application programming interface proved by Stackoverflow is used to extract the dataset.The language used in this project for back-end implementation is Python language because it provides a variety of libraries that are helpful for the types of data science projects. The frameworks used in the project to implement the back-end functionalities are Microsoft Visual Studio Code, Jupyter Notebook. and Microsoft Excel.

The front end design of this project is developed using REACT. It is the JavaScript library used for used for building user interfaces. Along with react, Bootstrap language is used for making the website fully responsive. Microsoft visual studio code is used as a code editor for building the react web application.

Chapter 6

System Testing and Evaluation

Testing is one of the most crucial and challenging phase and it ensures that the software product or anything related to it working effectively and correctly. It evaluates the software system and verifies that the software product does what it is supposed to do. Testing helps in improving system performance by figuring out the flaws and improvements for the system.

6.1 Software Testing Techniques

Software testing techniques are standard models used to ensure whether the system at hand is performing the functionality it was initially designed for, that is whether the system is fulfilling the objectives, functional and non-functional requirements defined in the early stages of the project. The testing of our Expert Ranking in Social Question Answering Forum was rigorous. Each module, technique and algorithm was checked thoroughly during and after development. Some of various testing techniques we used are given below:

- Graphical User Interface Testing
- Usability Testing
- Unit Testing
- System Testing
- Black Box Testing
- White Box Testing
- Acceptance Testing

6.2 Graphical User Interface Testing

The graphical user interface is the essential component of any system. it determines how the user will interact with the system and how the system will do when user performs certain operations.

In this section, the testing of our system's GUI has been described. We have used user interface testing and made ensure that our system is according to the requirements discussed in chapter 3. For this purpose, we have tested the main layout of the system. The layout is user friendly and clearly portrays the results. We have tested different topics by selecting them from drop-down list. The topics have successfully showed the desired results. We tested the expert authors by visualizing the results. The visualizations clearly shows that the result is according to the specified requirements.

6.3 Usability Testing

Since the purpose of usability testing is making human operators happy with the experience of using the interface. The usability testing is performed by our stakeholders and the test results showed that the system targets to the requirements we specified. The results provided by our system confirms that it can be integrated with any online discussion forum.

6.4 Unit Testing

We performed unit testing on back end module of the system such as topic modeling and link analysis. The system showed correct results on each module. The results generated by the system met all the requirements. The Expert Ranking in Social Question Answering Forum passed all the unit tests.

6.5 System Testing

System testing is performed by injecting results of the back end into the front end in an efficient way. We combined back-end and front-end modules and we then developed a website using ReactJS. Chapter 5 shows the efficient working of our system. Our system passed all system tests.

6.6 Black Box Testing

We checked the output results of our system concerning the functional specifications by manually comparing them from stackoverflow's website. Our system provided the results same as it provided at back-end also the results matched with the stackoverflow website.

6.7 White Box Testing

We internally checked the inner working of our system from user's (students, supervisors, target audience who frequently interacts) view to ensuring everything is working fine. The system passed white-box tests.

6.8 Acceptance Testing

In this type of testing, We have checked whether the actual requirements of the system are satisfying or not. We checked the functional requirements that were mentioned earlier in the objectives section whether they are achieved or not. The system passed all acceptance tests.

6.9 Test Case

We designed some test cases to test the executable functionality of Expert Ranking in Social Question Answering Forum System.

6.9.1 Test Case 1: Validating the Existence of Questioner

In this test case, the topic is selected from drop-down list and the list of questioners is displayed. The questioners are then checked manually from stackoverflow website whether they exists or not.

Table 6.1: Test Case Validating Questioner

Test Case	**01**
Description	To validate whether the questioners who have asked different questions actually exist or not. We validated each questioner by checking their IDs on stackoverflow website.
Initial State	The results must be shown on the screen which contains user IDs.
Functions to be tested	Validation successful or not
Test Execution	• Display the list of users by selecting any topic from drop-down menu. • Compare the user IDs on stackoverflow website.
Expected Result	The user should be found on the website.
Actual Result	The user is present on the website
Status	Pass

6.9.2 Test Case 2: Validating the Presence of Author

In this test case, the topic is selected from drop-down list and the expert/s of that certain topic is checked on stackoverflow website whether they exists or not.

Table 6.2: Test Case Validating Expert

Test Case	**02**
Description	To validate whether the authors who have replied to the questions actually exist or not. We validated each author by checking their IDs on stackoverflow website.
Initial State	The results must be shown on the screen which contains author IDs.
Functions to be tested	Validation successful or not
Test Execution	• Display the list of experts by selecting any topic from drop-down menu. • Compare the experts' IDs on stackoverflow website.
Expected Result	The expert should be found on the website.
Actual Result	The expert is present on the website
Status	Pass

6.9.3 Test Case 3: Topic Modeling

In this test case, we have verified whether the topic has been assigned to the particular question or not. The results show that every question has a certain topic which indicates the topic has been correctly modeled against the question.

Table 6.3: Test Case Topic Modeling

Test Case	**03**
Description	To identify whether the topic is assigned to each question or not.
Initial State	The browser should be started and questions should be listed.
Functions to be tested	Topic assignment successful or not
Test Execution	• The questions are trained for topic modeling after performing LDA. • The topic has been assigned to every question after the model is trained.
Expected Result	The topic should be assigned to every question.
Actual Result	The topic has been modeled successfully.
Status	Pass

6.9.4 Test Case 4: Verifying the Links of Authors

In this test case, we have identified whether our proposed algorithm works correctly or not. The proposed algorithm finds the connections of authors and ranks them accordingly.

Table 6.4: Test Case Link Analysis

Test Case	**04**
Description	To find the links of authors who answered certain threads along with the occurrence.
Initial State	The PageRank algorithm should be fed accordingly.
Functions to be tested	Link analysis successful or not
Test Execution	• The graph of authors along with threads is created. • The PageRank calculated the links of authors belong to certain threads.
Expected Result	The links of authors should be calculated.
Actual Result	The link analysis has been done successfully.
Status	Pass

6.9.5 Test Case 5: Finding the Experts

In this test case, we have found the expert/s of the particular topic. There can be more than one expert also. The authors have been ranked by the algorithm and the expert is found.

Table 6.5: Test Case Finding Experts

Test Case	**05**
Description	To rank the authors based on their connections and find the expert/s of particular topic.
Initial State	The PageRank algorithm should have links of each author.
Functions to be tested	Finding Expert successful or not
Test Execution	• PageRank algorithm have all the links. Assigning the PageRank value to the expert/s and discarding the least important authors.
Expected Result	The authors should be ranked correctly.
Actual Result	PageRank found the expert/s successfully.
Status	Pass

Chapter 7

Conclusions

In this project, we investigated the expertise level users in an online discussion forum and proposed an effective expert ranking algorithm. Most of the people need suggestions from the trusted experts for their questions in online discussion groups. The utility and commercial viability of this system lies in the fact that these discussion groups such as Facebook discussion group face the problem of ambiguous information. The proposed system makes sure that an expert is provided based on the social importance and the particular content information of thread the user is talking about. We have taken a small dataset for simplicity but the proposed results are clearly showing the effectiveness of the proposed system. As shown in the front end design of our project as an experiment, the content information of a thread allows us to find the most relevant or reputed users to certain topics/queries and the importance of users depends on their social networks. Then a ranking algorithm named ExpertRank had produced to determine the expertise level of users in this social network.

As a future direction, possible improvements might be looking for an approach to detect the scam users or bots who itself posts the question and answer the question. We may modify our ranking algorithm in different contexts and build expert network based on the publications of the experts and the citations belongs to them. Also, the tag-specific recommendation can also be included in this system where users are only recommended to the post which belongs to their tag and user's neighbours are also indulging. This will make ranking the expert to be more precise.

Appendix A

Data Dictionary

Table A.1: Data Dictionary

TERMS	DESCRIPTION
NLP	**Natural Language Processing** is the sub-field of artificial intelligence that allows computer program to understand human language when the interaction occurs. NLP provides the possibility to read text, hear sounds etc. to the computers.
SNA	**Social Network Analysis** primarily focus on relations among people and organizations. The interactions among the individuals is analyzed by using graph and network theory.
PageRank	**PageRank** is the algorithm that was proposed by Google to rank the websites according to their importance. In this project, the authors have been ranked instead of websites according to their importance on social networks.
WordNet Lemmatizer	**WordNet** is basically a large lexical database for the English language for creating the structural relationships among words. Here, **Lemmatizer** is used on text data because it understands the context of words and converts the words into their meaningful base form.
Stopwords	**Stopwords** does not add meaning to the sentence that's why they have been removed from the text. They are predefined and cannot be used to describe the topic of the content.
Stopwords	**Stopwords** does not add meaning to the sentence that's why they have been removed from the text. They are predefined and cannot be used to describe the topic of the content.
Topic Modeling	**Topic Modeling** is actually unsupervised machine learning algorithm that we used to get insights from large number of textual data. We discovered the extracted topics from multiple documents (documents mean the questions) and assigned the meaningful name to particular document.

LDA	**Latent Dirichlet Allocation** is used for topic modeling. The major benefit of using it is that it creates the topic per document and words per topic model. This makes it easy because it breaks the corpus into lower dimensional matrices.
REACT JS	**ReactJS** is JavaScript library that creates the web pages of highly dynamic and responsive to user input. We have created encapsulated components that supports their own state and then we have composed them to make complex user interface.
Victory Chart	**Victory** is actually data visualization library for React. it contains many components for visualizing the data. **Victory Charts** can show any data in the form of bar charts making it easy to understandable by the user.
Victory Pie	**Victory Pie** shows the dataset as a Pie. The data is divided into slices and each slice is displayed in the form of circle.

References

[1] Yang, Z., Liu, Q., Sun, B. and Zhao, X. (2019), "Expert recommendation in community question answering: a review and future direction", International Journal of Crowd Science, Vol. 3 No. 3, pp. 348-372.

[2] M. Bouguessa, B. Dumoulin, S. R. Wang et al., “Identifying authoritative actors in question-answering forums: The case of Yahoo! answers,” in Proceedings of the Knowledge Discovery and Data Mining, pp. 866–874, Las Vegas, NV, USA, August 2008.

[3] Tondulkar, R., Dubey, M., Desarkar, M. S. (2018). Get me the best: Predicting best answerers in community question answering sites. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 251–259). ACM.

[4] X. Cheng, S. Zhu, G. Chen and S. Su, "Exploiting User Feedback for Expert Finding in Community Question Answering," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 295-302, doi: 10.1109/ICDMW.2015.181.

[5] Zhao, Nan Cheng, Jia Chen, Nan Xiong, Fei Cheng, Peng. (2020). A Novel Expert Finding System for Community Question Answering. Complexity. 2020. 1-8. 10.1155/2020/5346085.

[6] J. Jiao, J. Yan, H. Zhao and W. Fan, "ExpertRank: An Expert User Ranking Algorithm in Online Communities," 2009 International Conference on New Trends in Information and Service Science, 2009, pp. 674-679, doi: 10.1109/NISS.2009.75.

[7] Roy P.K., Jain A., Ahmad Z., Singh J.P. (2021) Identifying Expert Users on Question Answering Sites. In: Goyal D., Bălaş V.E., Mukherjee A., Hugo C. de Albuquerque V., Gupta A.

[8] "StackOverflow API,"[online],Available: https://data.stackexchange.com/stackoverflow

[9] Roy P.K., Jain A., Ahmad Z., Singh J.P. (2021) Identifying Expert Users on Question Answering Sites. In: Goyal D., Bălaş V.E., Mukherjee A., Hugo C. de Albuquerque V., Gupta A.K. (eds) Information Management and Machine Intelligence. ICIMMI 2019. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-4936-6$_3$2

[10]$Husain, O. (2019). Expert findingsystems : Asystematicreview. MDPIAG.$

[11] Yang, Liu et al. “CQArank: jointly model topics and expertise in community question answering.” Proceedings of the 22nd ACM international conference on Information Knowledge Management (2013): n. pag.

[12] Shahriari, Mohsen, Sathvik Parekodi and Ralf Klamma. “Community-aware ranking algorithms for expert identification in question-answer forums.” Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business (2015): n. pag.

[13] Fu, Yupeng Xiang, Rongjing Liu, Yiqun Zhang, Min Ma, Shaoping. (2007). Finding Experts Using Social Network Analysis. 77-80. 10.1109/WI.2007.14.

[14] Kumar, Akshi Sangwan, Saurabh. (2018). Expert Finding in Community Question-Answering for Post Recommendation. International Journal of Engineering Technology. 7. 151. 10.14419/ijet.v7i3.4.16764.

[15] Faisal, M.S., Computers in Human Behavior (2018), https://doi.org/10.1016/j.chb.2018.06.013

[16] Sergio Jimenez, Fabio N Silva, George Dueñas, Alexander Gelbukh, ProficiencyRank: Automatically ranking expertise in online collaborative social networks, Information Sciences,Volume 588,2022,Pages 231-247,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2021.11.067. (https://www.sciencedirect.com/science/article/pii/S0020025521011890)