Predicting Best Answer in Community Question Answering System Using Naive Bayes Algorithm

Sayali Sonawane
Department of Information Technology Student, SCOE Pune, India sonawanesayali1999@gmail.com
Dr. K. S. Thakare
Department of Information Technology Professor, SCOE Pune, India kalpana.thakare_scoe@sinhgad.edu
Vaishnavi Kolte
Department of Information Technology Student, SCOE Pune, India vaishnavikolte6@gmail.com
Ashwini Dhavare
Department of Information Technology Student, SCOE Pune, India ashwinidhavare.ad@gmail.com
Pranita Jejurkar
Department of Information Technology Student, SCOE Pune, India pranitajejurkar14@gmail.com

Article Info
Volume 83
Page Number: 2000 - 2009
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 18 March 2020*

Abstract

Community question answering sites are eminent online community that has bring along users to another milestone of knowledge provision to let the users ask queries. There are number of increasing questions that get posted but may not get fixed in short amount of time as domain expert may not find questions that he/she is capable of answering and also finding the best answer among multiple answers is another challenge. Hence, we propose a new technique to question routing system using text classification Naive Bayes algorithm and Natural Language Processing technique. This system provides technical and non-technical communities both. Every community have experts provided, which will answer the questions routed to them. Propose system mainly works on ranking of answers and ratings given by user to find the best answer. Ranking is achieved through ratings given by the users. In some of conditions unsatisfied answers on system get resolve by online forum with direct communication between the expert and user.

Keywords – *Community Question Answering, Question Classification, Answer ranking, Natural language processing (NLP).*

I. INTRODUCTION

Question Answering Services is a new area of study in the field of Information Retrieval (IR). Community question answering systems are prototype like forums. On the forum people share their views, opinions and also ask questions to clear their doubts. And these community question answering systems are used all over the globe, various questions can be found along with their answers. User can find solution to his/her problem this system. Question Answering (QA) websites Such as Stack Overflow, Answers.com and Quora is gaining popularity, because of the flexibility of these websites which try to provide information like answers of asked questions or related answers of asked questions which will help user. Question Answering Systems (QASs) used earlier were domain restricted and had limited capabilities in providing answers to user. Frequently asked Q&A s must be categorized into different communities depending on the questions which is being asked by the users, most discussed data sources and different forms of answers generated. Since large number of QASs has been developed, research in the domain of QAs has begun. Identifying the future scope of research is a fundamental way of arising the survey of QASs. This survey gives an outline of current QASs, its system structure and suggests the future scope of the research. There are many community question answering systems which are useful for people for the searching question of their interest and getting their answer on the web forums but every time user searches new question in return the user gets a lots of answers. The analysis of those answer is time consuming and laborious. User cannot find which answer is best answer. Proposed system work on the rank model based on QA pair rating and online forum support. When the user finds the answer as exactly, he wants; user gives rating to that answer and considers answerer as an expert. And next time the preferences are given to that expert for good answers and accordingly his reputation increases in community. When user asks question, which is already existing in the system, that time system provides previous answer which has the highest rating and that answer can be consider as best answer. Support of forum is provided for user assistance, using which user can directly communicate with experts if he satisfied with answer of expert.

II. Literature Survey

The joint implicit and explicit neural network (JIE-NN) models are applied for online question recommendation in CQA. Textual content and social connections into an end-to-end neural network are a heterogeneous information source which maintains flexibility of proposed system. To extract latent textual features, Convolutional Neural Networks (CNNs) is used to measure the similarities between the given questions. For question recommendation purpose they have used implicit user groups. It repeats the back-propagation process of neural network for dynamically clustering of user group. The main advantage of the system is that it involves experts in the system and Community wise question routing is done. Ranking method is not used for ranking answers.[1]

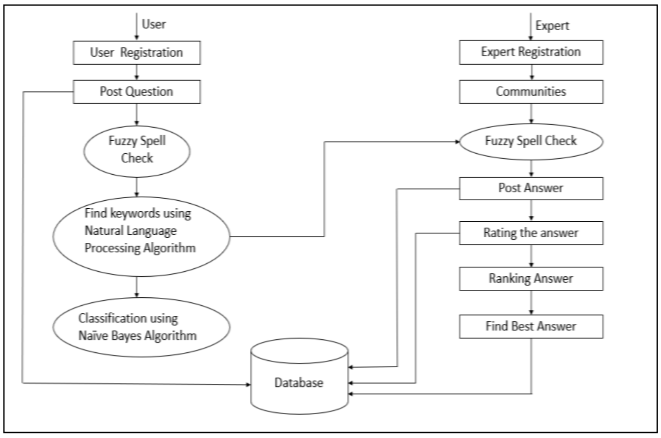
Tirath Prasad Sahu, at el has considered the difficulty of question recommendation as a classification task. They developed a variety of local and global features which seize different aspects of questions. Local feature introduced here which are needed to access the local information about question and user history. This group includes features about question such as the length of the title (subject), length of the detail, and 5WH question type (why, what, where, when, who). In global feature group it includes category level features such as average title length and average detail length.[2]

The correctness of the system can be improved by using quality prediction features, ranking scores for best answer choosing, text feature extraction and text categorization. The qualities of the answer for user are improved by using the Best answer choosing method. But the drawback such as quality of response is uncertain and the category system is not clear and detailed are present in the system. [3]

The main parts of this research are Natural Language Processing and text mining techniques. The Sentiment Analysis is used on comments and spell checking for answers. The accepted answer is represented in organized text format such as having less spell errors in the system, but Self-answering feature is not provided. [4]

They have demonstrated with information retrieval techniques. There are different ways of creating profiles for finding experts in community-based question answering services. The question posted by the user can be treated as query and expert profiles can be treated as records. By using language models these profiles are ranked which represents information retrieval techniques. The language models use in this work are: the query likelihood model, the relevance model, and the cluster-based language model. User who have higher ranked profile can be considered as experts for answering the given questions.[5]

Proposed an approach to obtain question and answer from a question answering services that can be provided as a valuable resource to drill retrieval models. It can recognize similar questions and users can directly acquire the answers.[6]



III. Proposed System

Fig. 1: Community Question Answering System architecture

Question Answering system represent the research in the domain of social networks, information retrieval and knowledge management systems. The newly proposed system includes all the features of existing system and recovers its drawbacks. It involves community experts for answering queries and those experts belong to different communities. Experts can also share their study material to users. By using this resource material user can get more knowledge. There are different communities involved such as technical (Java) and non-technical (Music and Sports). Question gets classified according to these communities. Only Experts can give the answers. Many people have already asked same information before, is provided by the system itself. Users can give ratings to the answer.

Fig. 1 shows the details of the proposed system, in which the user is the main person related to the system, he/she get registered themselves with all the information with his interest, education and knowledge. The proposed model is divided into three phases.

Phase I: User Registration and posting a question

User will register in the system by entering his details. User has authority to post a new question. While posting a new question the user can make spelling mistakes or may not know the spelling of many words therefore a fuzzy spell checker API, which is basically a spelling error corrector is used. After posting a question, system will find keywords from the questions using NLP algorithm. This algorithm will compare these keywords with the available dataset of different communities. The output of this NLP algorithm is further forwarded as an input to the Naive Bayes algorithm which is used for question distribution in the respective community.

Phase II: Expert Registration and posting answer

In this registration process user will enter his details and select the community according to his domain of expertise. This user has to undergo an online test in the prescribed domain. After successful completion of the test as per the given criteria, user can become an expert in that community of the system. User will be now eligible as an expert and will be able to solve the queries ask by different users. The question ask by the user and answer posted by the expert will get stored in the database.

Phase III: Ranking and finding best answer

In this phase user can give rating to the answer when he gets satisfied with the solution provided by the expert. To show the accepted answers quickly which is being asked by the users could specify as the best or accepted answers. According to rating given by the user, answer will be ranked and the highest rating answer will be on the top of the list. Depending on ratings, rewards will be given to the expert for his answer and the reputation of the expert will increase accordingly in the community.

IV. UML Designs of System

1. Use Case Diagram:

A use case diagram is a pictorial representation of a user's interrelation with the system. Use cases are meant for specification of the interaction between the systems itself. The end users are also involved in the system which is called as actors. There is a responsibility of each actor for particular task in the system. Use case must have unique name and that name should describe overall functionality of use case.

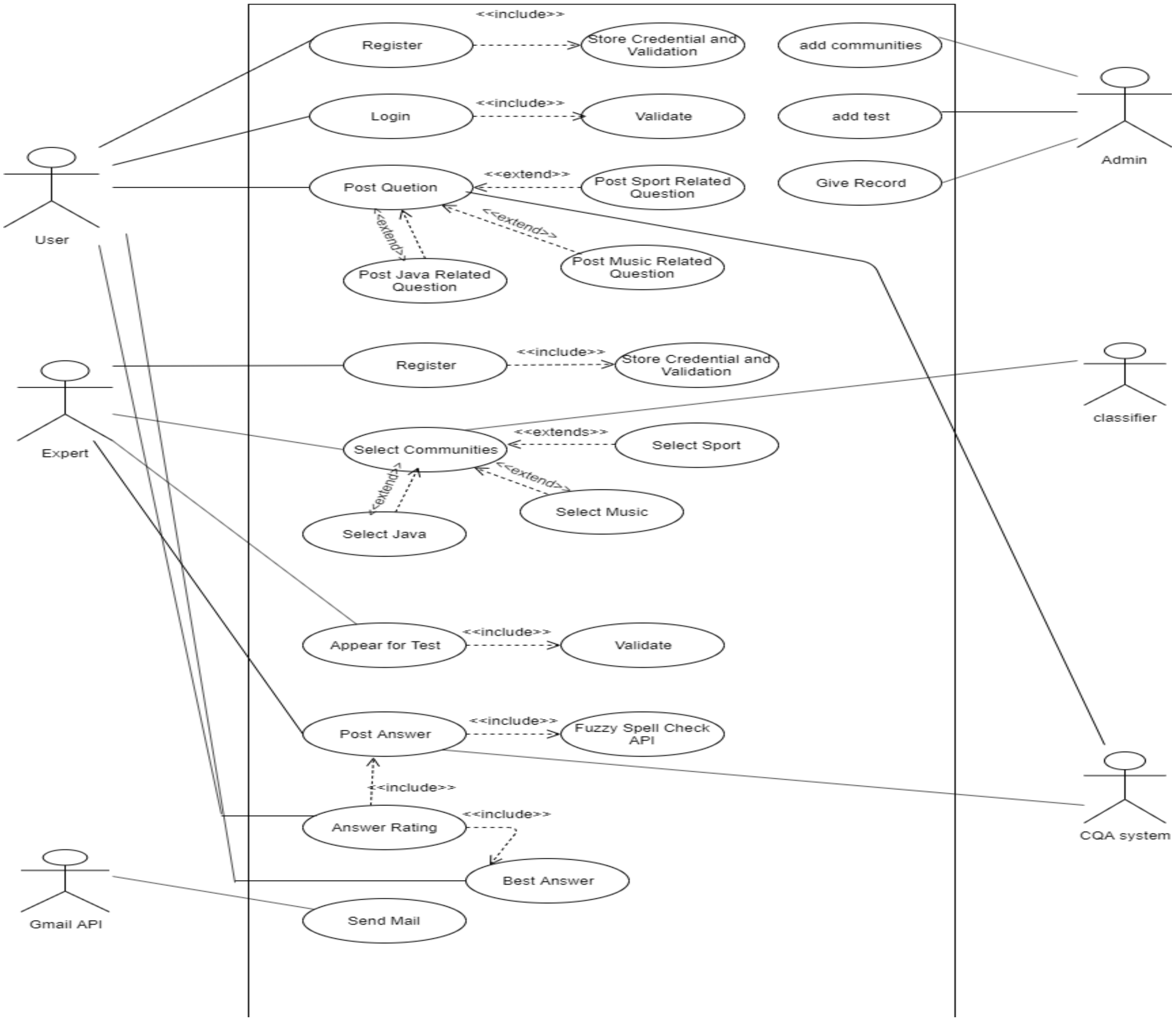


Fig. 2: Use Case Diagram

2. Sequence Diagram:

Sequence diagram is a type of interaction diagram. It shows how objects communicate with each other and in what order. Sequence diagram mainly uses the object timeline for time ordering of messages. Objects in the sequence diagram are the instances of elements like nodes, components.

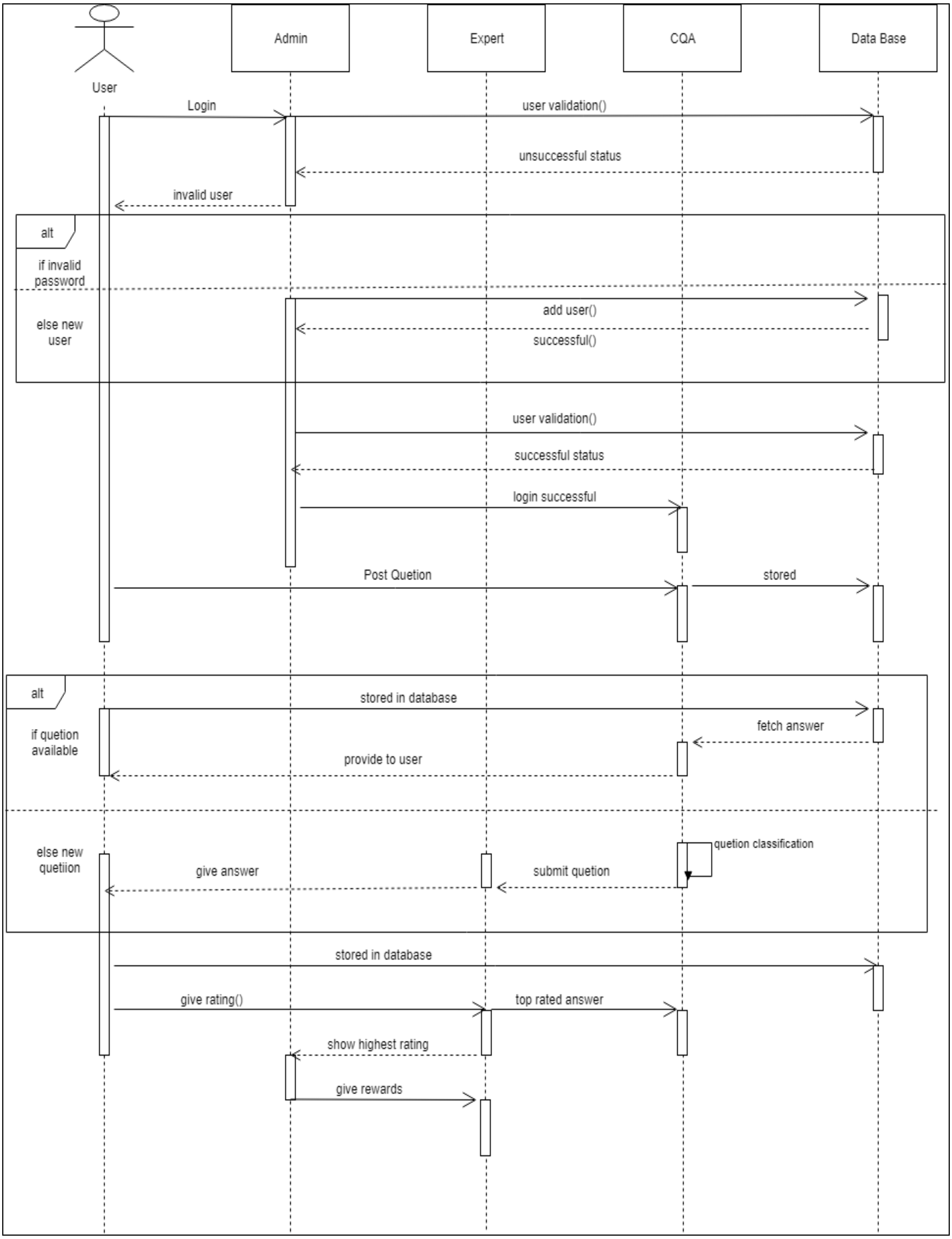


Fig. 3: Sequence Diagram

3. Class Diagram:

Class diagram is a static diagram. It is used to represent the latent view of an application and collection of classes, interfaces, their interrelationships, collaboration of classes. It describes how things are well organized. A group of class diagrams represents the whole system and it is also called as the foundation for component and deployment diagrams.

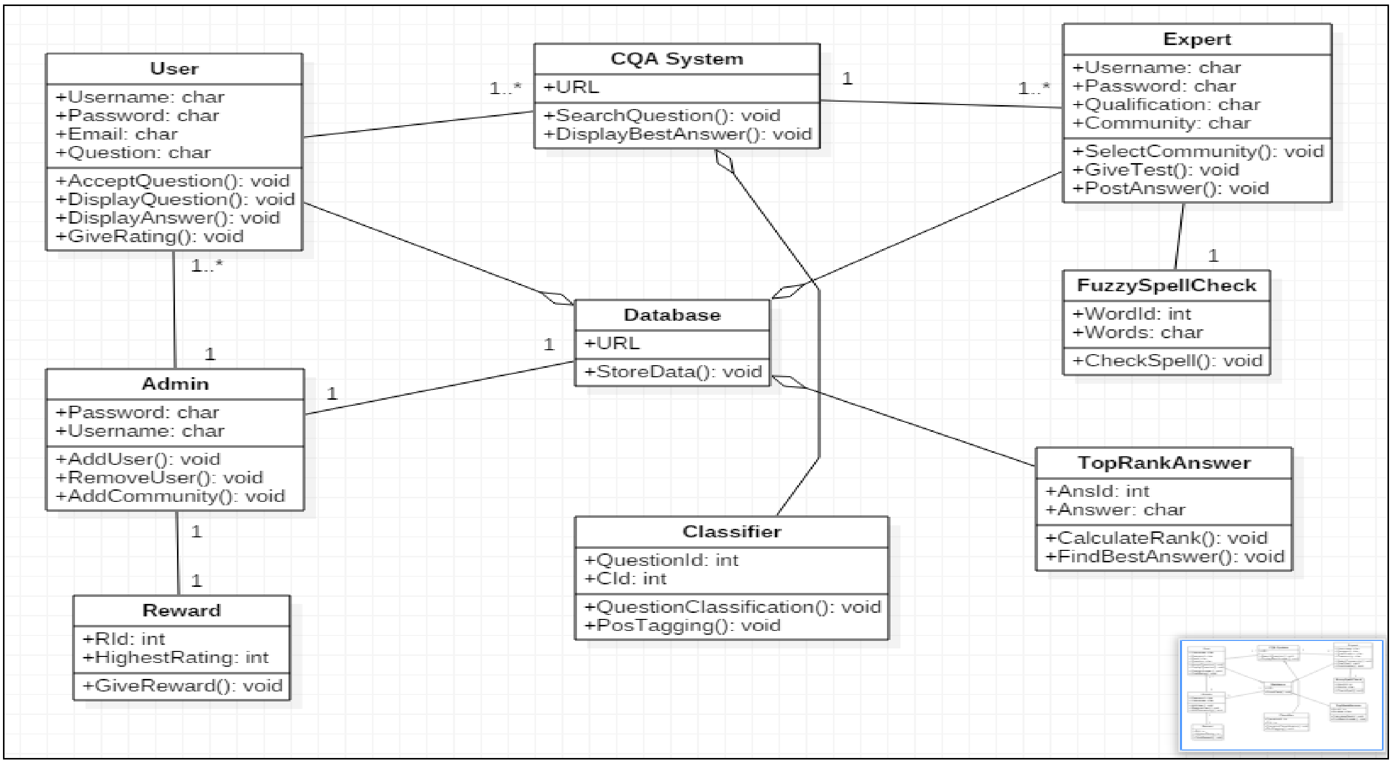


Fig. 4: Class Diagram

V. Implementation And Algorithm

Important algorithms used for the system are explained in this section.

1. Natural Language processing (NLP)

It is a technology that is used to alter the unorganized text in documents, suitable for analysis or to drive data mining algorithms. Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering concerned with the interactions between computers and human (natural) languages. It conveys how to program computers to process and analyze large amounts of natural language data. In the proposed system the input to this algorithm is question posted by the user. NLP algorithm will find out the keywords from the question which will help to classify the question according to the community. Working of NLP algorithm is as follows:

Step 1: Segmentation of Sentence

This is the process that separates the paragraph or text document into single sentence.

Step 2: Tokenization of Word

Tokenization is used to separate text into small units such as sentences or words. Unstructured text can be converted into structured format.

Step 3: Identifying token of the Part of Speech

This is the process that deals with the token of the text and it further specifies the categories of tokens like noun, verb, adjective, etc.

Step 4: Text Lemmatization

The morphological and structural analyses of the words are taken into consideration. It normally aims to remove the inflectional endings and consider the dictionary form of the word.

Step 5: Stop word Identification

Stop words are the words that are commonly appear in any text. We can filter these words as they don't tell much about data.

Step 6: Dependency parsing

It will identify how all the words in sentence are correlated to each other.

In this process a tree of noun phrases is built and that assigns a single parent to each word in the sentence.

Step 7: Named Entity Recognition (NER)

NER represent real-world concepts by detecting and labeling the noun. They use the context of how words are represented in the sentence and it detects which type of noun word appears.

Step 8: Coreference Resolution

Coreference resolution is the function of finding all expressions mentioned in text that refer to the same real-world entities.

2. Naive Bayes Algorithm:

For text classification purpose Naive Bayes algorithm is used. Naive Bayes classifier works on large datasets with high accuracy and speed, it assumes that the impact of a particular feature in a class is independent of other features. After finding the keywords through NLP algorithm, keywords will match with the dataset and find out that question belongs to which community. And then question will get classified according to the community. Naive Bayes algorithm works as follows:

Step 1: Read the Training Dataset.

Step 2: Convert the Training dataset into Frequency Table.

Step 3: By finding the probabilities create likelihood table.

Step 4: The posterior probability of each class is calculated using Naive Bayes equation.

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)}$$

Where, P(c|x) = Posterior Probability

P(x|c) = Likelihood Probability

P(c) = Class Prior Probability

P(x) = Predictor Prior Probability

Step 5: The class which has maximum posterior probability will be the result of prediction.

Posterior probability: P(c|x) is a posteriori probability of x, i.e. probability of event after outcome is occurred.

Likelihood Probability: It is an infinite set of possible probabilities, given an outcome.

Prior Probability: It is the probability of an event before evidence is seen.

VI. Results And Analysis

Fig. 5 is the User Interface of the proposed method. Here Admin can login to the system and user can select the community.

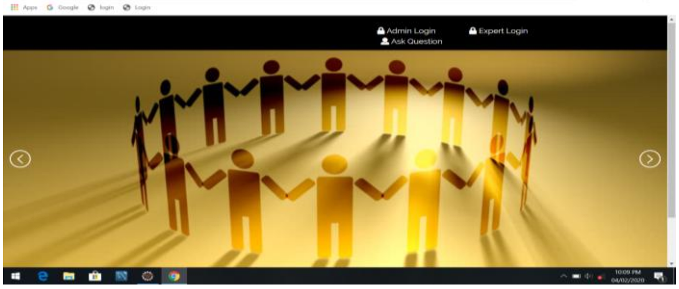


Fig. 5: Home Page

Fig.6 shows admin login portal. System will authenticate admin's identity by accepting credentials. Here admin can manage all the user and expert activities. Admin is also able to give reward to the best expert according to their performance in the community.

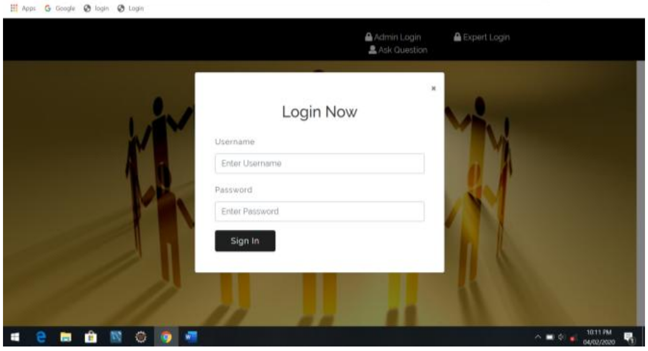


Fig. 6: Admin Login Portal

Fig.7 represents admin page. Admin can add keywords which are unknown to the system. Manually admin can add keywords for particular community.

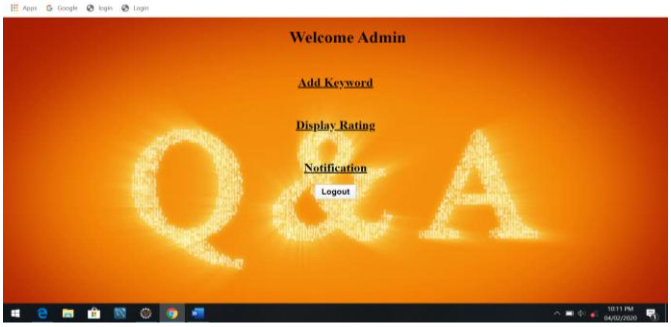


Fig. 7: Admin Page

Admin can add new keywords manually. For this admin need to select community and type a keyword and register it in the system as shown in fig.8.

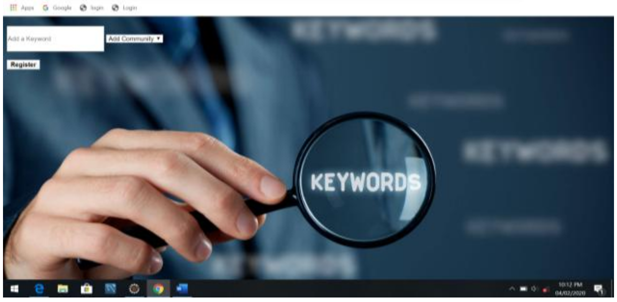


Fig.8 : Window to add keyword

User first need to choose a community and then can enter new keyword as per the community. After clicking on “Register” button keyword will get added to the database and successful message will get displayed on the scree

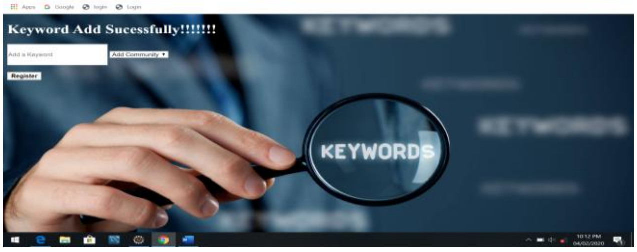

Keyword Add Sucessfully!!!!!!!

Fig. 9: Window displaying Keyword added successfully message

Fig. 10 shows Expert Registration Form where expert can register by entering his login details and domain of interest.

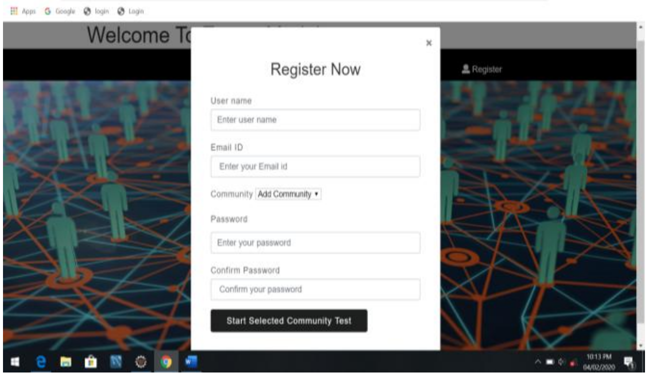
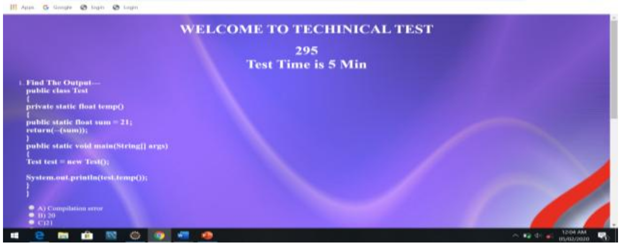


Fig. 10: Expert Registration Form

Fig.11 shows the expert test page, where expert need to qualify the test to become an authorized expert.


WELCOME TO TECHINICAL TEST

295

Test Time is 5 Min

1. Find The Output—

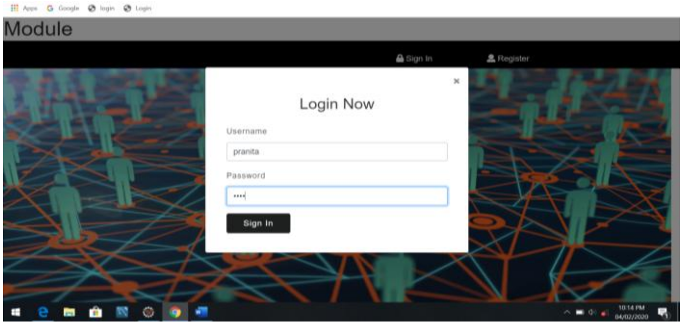
```
public class Test
{
private static float temp()
{
public static float sum = 21;
return(--(sum));
}
public static void main(String[] args)
{
Test test = new Test();

System.out.println(test.temp());
}
}
```

- ○ A) Compilation error
- ○ B) 20
- ○ C) 21

Fig. 11: Expert Test Window

Fig. 12 displays expert login page by which expert can login to the system.



Login Now

Username: pranita

Password: ····

Fig. 12: Expert Login Form

Fig. 13 shows window of Java Community, where user can see questions related to Java Community and also can ask questions.

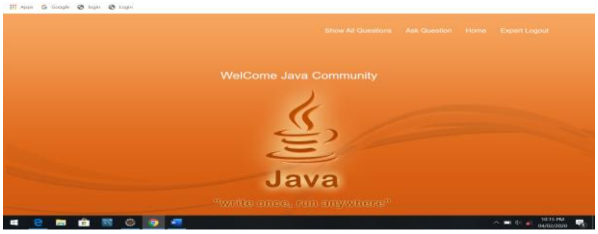
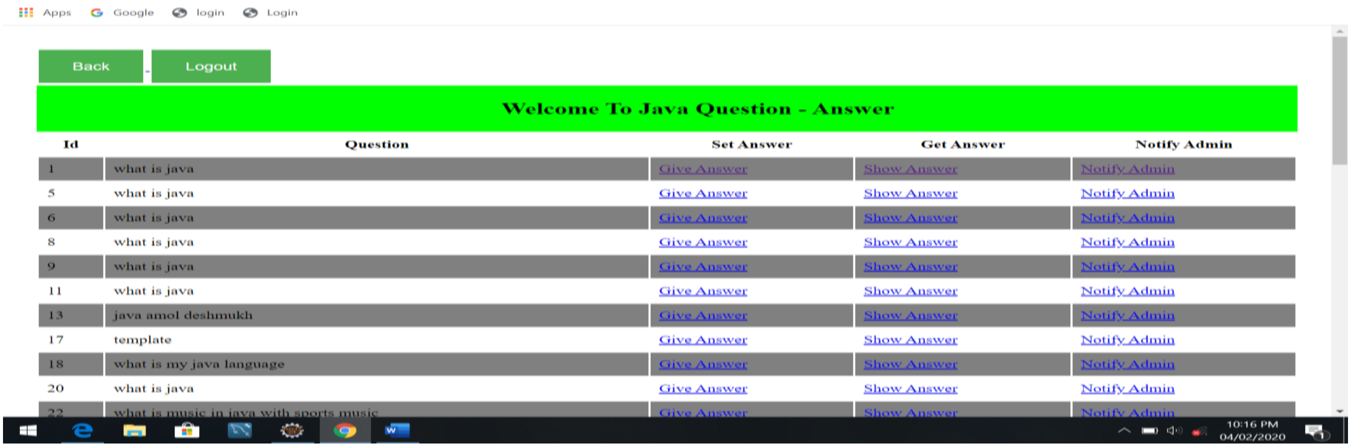


Fig. 13: Window of Java Community

Fig.14 shows all questions of Java Community. User can see the answers, Expert can give answer and if any query occurs user can notify to the admin.

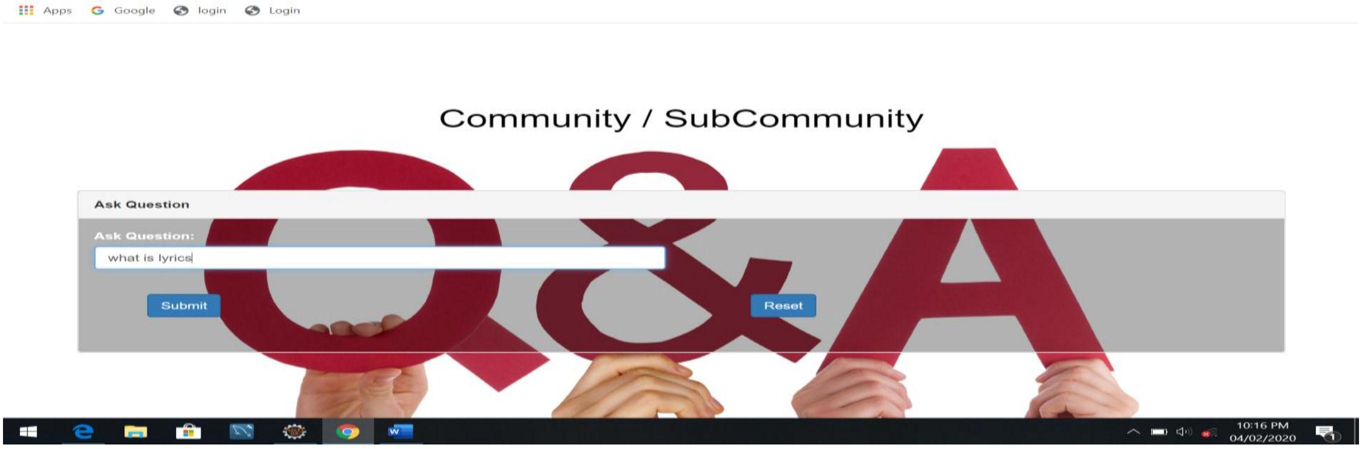


Welcome To Java Question - Answer

Id	Question	Set Answer	Get Answer	Notify Admin
1	what is java	Give Answer	Show Answer	Notify Admin
5	what is java	Give Answer	Show Answer	Notify Admin
6	what is java	Give Answer	Show Answer	Notify Admin
8	what is java	Give Answer	Show Answer	Notify Admin
9	what is java	Give Answer	Show Answer	Notify Admin
11	what is java	Give Answer	Show Answer	Notify Admin
13	java amol deshmukh	Give Answer	Show Answer	Notify Admin
17	template	Give Answer	Show Answer	Notify Admin
18	what is my java language	Give Answer	Show Answer	Notify Admin
20	what is java	Give Answer	Show Answer	Notify Admin
22	what is music in java with sports music	Give Answer	Show Answer	Notify Admin

Fig. 14: Java Question Answer Window

In fig.15, user can ask the question.



Community / SubCommunity

Ask Question

Ask Question: what is lyrics

Fig. 15: Window for asking question

VII. CONCLUSION

A centralized community question answer system is proposed which maintains group of users in the form of communities according to their interests. This helps in distributing or circulating the question posted by user to the person or people who acquire adequate amount of knowledge regarding the subject (expert). To identify the experts of given subject, the system uses test for particular subject. Support of forum is provided for user's assistance, using which user can directly communicate with the expert if he satisfies with the answer of expert.

VIII. REFERENCES

[1] Hongkui Tu, Jiahui Wen, Aixin Sun, Xiaodong Wang, "Joint Implicit and Explicit Neural Networks for Question Recommendation in CQA Services", National Institute of Defence Technology, pp.73081-73092, vol.6, 2018.

[2] Tirath Prasad Sahu, Naresh Kumar Nagwani, Shrish Varma, "Selecting best answer: An Empirical Analysis on Community Question Answering Sites", International Journal of National Institute of Technology, pp.47974808, vol.4,2016.

[3] Zive Zhu, Xiaoqian Liu, Huakng Li, Tao Li, "UB-CQA: A User Attribute Based Community Question Answering System", In proceeding of International Conference on Intelligent System and Knowledge Engineering.

[4] Fatemeh Eskandari, Hamir Shayestehmanesh, Sattar Hashemi, "Predicting Best Answer Using Sentiment Analysis in Community Question Answering System", International journal of Amirkabir University of Technology, pp.5357,2015.

[5] Xiaoyong Liu, Matthew Koll. "Finding Expert in Community-Based Question Answering Services", Center for Intelligent Information Retrieval, Computer Science Department University of Massachusetts, Amherst, MA 01003,2005.

[6] Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee. "Finding Similar Question in Large Question and Answer Archives", Center for Intelligent Information Retrival, Computer Science Department University of Massachusetts, Amherst, 2005.

[7] Baichuan Li, Iewin King, "Routing Question to Appropriate Answers in Community Question Answering Services", Department of Computer Science and Engineering The Chinese University of Hong Kong. pp1585-1588,2010.

[8] Ivan Srba, Marek Grznar, Maria Bielikova, "Utilizing Non-QA Data to Improve Question Routing for Users with Low QA Activity in CQA", In proceeding of International Conference on Advanced in Social Networks Analysis and Mining, pp129-136,2015.

[9] M. liu, Y. Liu, and Q. Yang, "Predicting Best Answer for new Question in Community Question Answering", in Proc. of the 11th Int. Conf. on Web-age Inf. Management, pp127138,2010.