EMPIRICAL ANALYSIS OF SUPERVISED MACHINE LEARNING TECHNIQUES FOR EXPERT MINING IN CQA SYSTEM

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

Dhiraj Kumar Gupta
2K16/SWE/06

Under the Supervision of

Dr. AKSHI KUMAR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

JUNE, 2018

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Dhiraj Kumar Gupta, Roll No. 2K16/SWE/06 student of M.Tech (Software Engineering), hereby declare that the project Dissertation titled "Empirical Analysis of Supervised Machine Learning Techniques for Expert Mining in CQA System" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **DHIRAJ KUMAR GUPTA**

Date:

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the project dissertation titled "Empirical Analysis of Supervised Machine Learning Techniques for Expert Mining in CQA System" which is submitted by Dhiraj Kumar Gupta, Roll No. 2K16/SWE/06, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master in Technology (Software Engineering), is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for ant Degree or Diploma to this University or elsewhere.

Place: Delhi **(Dr. AKSHI KUMAR)**

Date: **SUPERVISOR**

Assistant Professor

Department of Computer Engineering

Delhi Technological University

ABSTRACT

Today's era is influenced by the social media indispensably. Apart from sharing, discussing and communicating online, people use Community Question Answering online services for finding answers to their queries and questions. Huge amount of information and contents has been accumulated in Community Question Answering platform, where the vital concern is finding of an expert who could respond to user's questions efficiently and accurately. In this paper, we attempt to search, study, examine and analyze the previous studies that have been used for finding experts in Community Question Answering portals. We intend to infer the research gaps pertaining to expert mining in Community Question Answering platforms in the past decade. The contribution of this work is significant as it will certainly aid the future researchers and practitioners in understanding the applications of expert mining in Community Question Answering.

ACKNOWLEDGEMENT

I am most thankful to my family for constantly encouraging me and giving me time and unconditional support while pursuing this research.

I am extremely grateful to Dr. Akshi Kumar, Asst. Professor, Delhi Technological University, Delhi, India for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for her extensive support and encouragement.

I also convey my heartfelt gratitude to all the research scholars of the Web Research Group (WRG) for their valuable suggestions and helpful discussions throughout the course of this research work.

DHIRAJ KUMAR GUPTA

(2K16/SWE/06)

TABLE OF CONTENTS

Candidate’s Declaration..ii

Certificate.. iv

Abstract..iv

Acknowledgement...v

Table of Contents..vi

List of Figures..viii

List of Tables..ix

List of Acronyms..x

Chapter 1 Introduction and Outline..1

1.1 Introduction...1

1.2 Research Objectives...3

1.3 Proposed Model...4

Chapter 2 Literature Review...6

2.1 Expert Finding in CQ.. 6

2.2 Perspective of Expert Mining in CQA...12

2.3 Approach towards Expert Finding in CQA..14

Chapter 3 Proposed Model...17

3.1 The Proposed Model...17

3.1.1 Dataset...18

3.1.2 Preprocessing and Features Extraction...20

3.1.3 Evaluation Criteria...21

3.1.4 Learning Model...21

3.1.5 Evaluation..22

Chapter 4 Experimental Results and Analysis...24

Chapter 5 Conclusion..27

Bilbliography..28

APPENDIX...31

LIST OF FIGURES

Figure 1.1. Different types of CQA sites...2

Figure 2.1. Perspectives of CQA...13

Figure 2.2. Representation of Graph based approach...14

Figure 2.3. Representation of Featured Based approach...15

Figure 2.4. Various methods used for Expert Finding...16

Figure 3.1. The predictive learning model for the CQA system...................................18

Figure 3.2. Dataset before preprocessed..20

Figure 4.1. Accuracy...24

Figure 4.2. Precision... 25

Figure 4.3. Recall...25

Figure 4.4. F-measure..26

LIST OF TABLES

Table 2.1 Summary of Studies Related to Expert Mining in CQA..9

Table 3.1Snippet of the Dataset Used...19

Table 3.2 Supervised Learning Technique Used...21

Table 3.3 Efficacy Measures used...22

Table 4.1 Performance measure results for different ML Techniques...24

LIST OF ACRONYMS

A	Accuracy
AUC	Area Under the ROC Curve
AP	Average Precision
AR	Average Rank
AS	Average Score
BAC	Best Answer Coverage
CA	Concordance Analysis
CDR	Cumulative Distribution of Rank
CQA	Community Question Answering
DCG	Discount cumulative gain
F	F-measure
GRMC-AGM	Graph regularized matrix completion advanced proximal gradient search method
GRMC-EGM	Graph regularized matrix completion expanded gradient method
KPI	Key performance indicator
K-NN	K-Nearest Neighbor
LR	Linear Regression
LSTM	Long Short Term Memory
M	Mean
MAP	Mean Average Precision
ML	Machine Learning
MRR	Mean Reciprocal Rank
MS	Median Score
n-DCG	Normalized Discount cumulative gain
NB	Naïve Bayes
NLP	Natural Language Processing
P	Precision
P@1	Precision for Top 1 user

P@5	Precision for Top 5 user
P@10	Precision for Top 10 user
PIAP	Point Interpolated Average Precision
QLL	query likelihood language
R	Recall
RR	Reciprocal Rank
RF	Random Forest
S	Success
SE	Standard Error
SVM	Support Vector Machine

Chapter 1 Introduction and Outline

This chapter briefly introduces the research work proposed in the thesis. Section 1.1 gives an overview of the research undertaken. Section 1.2 sets out the research objectives. Section 1.3 illustrates the proposed framework and the main contributions arising from the work undertaken. Finally, Section 1.4 presents an outline of this thesis describing the organization of the remaining chapters.

1.1. Introduction

Expert Mining is a difficult task that consider finding and identification of actual experts with their expertise in the particular domains. The phenomena is quite intricate as it involves the inquisitors who may ask questions based on incomplete and improper articulated requirements. Web 2.0 [1] has changed the way of thinking and expressing one's opinions and views. After the emergence of web 2.0 [2], websites have become more usable, securable and interoperable (work in other products, system and devices). People can gather or share their knowledge, experience, thought, queries from various knowledge sharing system, such as newsgroups, wikis, e-commerce, blogs, media, and bulletin boards. Ideas, thoughts and information are being shared and discussed on social media platform like Facebook and Twitter, e-commerce websites like Flipkart and Amazon, media websites like India times, NDtv and Times of India, Community Question Answering (CQA) websites like Quora [3], Stack Overflow and Yahoo! Answering. Among them, CQA websites are becoming more popular in terms of knowledge sharing platform. CQA services are communication platforms where people can exchange knowledge, fact, information and skill with each other based on question and answers. People can directly ask their questions or can search answers related to their questions asked by other users in the CQA websites.

CQA system helps in building a 'knowledge repository' that could be useful for all the users to obtain answers for their queries. People can ask queries of any domain on the CQA system. The prime objective of the CQA websites is to route the questions to an "Expert", who has the ability to provide correct and good quality answers. Then other user gives the "Up Vote" or "Down

Vote" according to the truth-ness, correctness, exactness and quality of the answers. The count of "Up Vote" or "Down Vote" indicates the extent to which the user is knowledgeable. Thus users can obtain satisfactory answers of good quality of their questions and save huge amount of time and resources. Yahoo! Answering, Stack Overflow, Quora, etc. are the most popular websites based of CQA. Fig 1.1 describes some most popular sites of CQA.

Fig 1.1 Different types of CQA sites

In our day to day life, CQA services play a very crucial role for finding the answers to the questions that are not easily available on the internet. If a user tries to search an answer for a question in the traditional way, then the time and cost of finding a solution for the query would be large. Whereas with the help of CQA services, a user will be able to find a number of solutions from all over the world of the posted query within a few minutes. People can ask queries of any domain provided it belong to at least one of the tag of the CQA system. For instance, Stack Overflow is concerned only with the domain of computer science and applications whereas Quora considers a wide range of domains like art, lifestyle, commerce, computer science, etc.

CQA services builds a very strong community network among its users. As every coin has its two sides, the CQA services have some negative aspects too and one cannot oversee its negative effects. Though CQA services have a very large community network, a user might have to wait for an answer for a long period due to following reasons (i) wait for the users, who answers to the question (ii) the answer may be incorrect, obnoxious or spam (iii) or use the archives of CQA sites. These archives often contain restricted answer sets and the user has to deal with the word-match constraint between his formulated question and archived questions. Thus Finding an expert in CQA system is necessary. Where an "Expert" can be defined as someone who is preferred to provide and present "high quality answers" to the queries of other users [4].

In the thesis, users will be classified into two categories i.e. experts and non-experts. Various machine learning algorithms are applied for the classification. The remainder of this chapter sets out the research objectives, describes the main contributions of the research work, and presents an outline of this thesis.

1.2. Research Objectives

Statement of Research Question

"Can we find a user who could be an expert in future using machine learning?"

Many a times, it may occur that inquisitors may not get good quality, appropriate, unambiguous and correct answers due to lack of availability of "Potential Experts" in that particular area. Potential experts are the specialists that have the knowledge and capability of becoming proficient in the future. Thus, this unifying research question can be broken down into the following two questions, each of which will be addressed by this research:

- What is the need to find the expert?
- Which supervised machine learning techniques are best to classify expert from non-expert?

Consequently, the three main research objectives of the work undertaken are:

I. **Research Objective I** – What are the most commonly used datasets in CQA for determination of expert?

II. **Research Objective II** – Which parameters or attributes can be used for finding experts within CQA systems?

III. **Research Objective III** – What is the scope of finding experts in CQA and its relevance in the current era?

The objective of this thesis is to propose a model to find the potential users who will become expert in future and then to apply various machine learning techniques to find the accuracy (A), precision (P), recall (R) and f-measure (F).

1.3. Proposed Model

The proposed model consists of two modules i.e., firstly labeling the dataset into experts and non-experts and secondly applying various supervised machine learning techniques to evaluate the accuracy of the proposed classifying model. The aim of classifying the users into expert and non-expert is to route the question to the expert so that a quality answer which is unambiguous, correct and complete is answered as a solution of the asked query. To determine the accuracy of the proposed model and to determine which supervised machine learning algorithm is best suited for our proposed model, various supervised machine learning models have been applied on the dataset. For the experimental purpose, Stack Overflow data is used as a dataset. Also, the various supervised machine learning algorithm used in the proposed model are Naïve Bayesian, Support Vector Machine, Random Forest, k-Nearest Neighbors and K-STAR.

1.4. Organization of Thesis

This thesis is structured into 5 chapters followed by references.

Chapter 1 presents the research problem, research objectives, justifies the need for and outlines the main contributions arising from the work undertaken.

Chapter 2 provides the essential background and context for this thesis and provides a complete justification for the research work described in this thesis.

Chapter 3 provides the details of the methodology employed and outlines of the proposed model that constitutes the proposed approach of the research.

Chapter 4 describes the experimental results obtained by applying the proposed model. It also presents the analysis to account for the tests performed.

Chapter 5 presents future research avenues and conclusions based on the contributions made by this thesis.

Chapter 2 Literature Review

This chapter discusses the background work in the research domains of expert finding on CQA systems. We present a state-of-art review of expert finding on CQA. The research gaps have been identified as issues and challenges within the domain which make it an active and dynamic area of research.

2.1 Expert Finding in CQA

In present, a very large amount of data is being shared by using web 2.0. There are numerous applications that are in use for sharing the data like weblogs, newsletters, microblogs, social media, wikis, CQA websites and many others on the internet. So far, numerous weblogs exist on the internet and the number of them is growing fast. On Wikipedia, volunteers from all around the world update over 3 million articles daily. Yahoo! Answers enticed about 60 million distinctive users and 160 million responses within one year of its launch [2]. In addition, there are more than 13 thousands users on java online community. Prominent companies like IBM, Google, Facebook, Microsoft and many others are using this high volume of Information and are making millions of profit by extracting useful data using data mining and business intelligence tools. In recent times, expert finding for CQA systems have grabbed much attention of the researchers but 15 years ago, expert finding was one of the most difficult task for the researchers. So far, two of the most important issues in these online environments are studying users for user modelling and expert finding. The first problem is routing question to an expert so that inquisitors or other users will get correct knowledge, justification, source as well as good quality of answers and the second problem is to identify the best answers from all answers. There are numerous models that have been proposed by the researchers for finding the expert in the CQA systems.

Literature has lot of studies illustrating expert mining in CQA and various researchers have proposed several models. One such model was given by the author Mohamed Bouguessa et al. (2008) [5], who proposed a Probabilistic model in ordered to choose an Authoritative User in

Question Answer communities. They used Yahoo Answering as the dataset. Another author Yao Lu et al. (2009) [6] had proposed a Latent Link analysis approach for finding the expert in the CQA services. They had compared the direct and the latent relationship links in the graph model. Experimental results showed that their approach performed better than the tradition link analysis model in terms of finding the expert in CQA. They had used Yahoo Answering as a dataset and had evaluated the system using Precision (P) as a performance parameter. Chirag Shah et al. (2010) [7] had proposed Logistic Regression model to evaluate and predict quality of an answer in the CQA. For evaluation of their model for Yahoo Answering repository, they had employed Mean (M), Standard Error (SE) as performance indicators. Aditya Pal et al. (2010) [8] had used Machine Learning Model to find the expert in the CQA systems. They had worked on criteria's like Precision (P), Recall (R), F1 Score (F1). Aditya Pal et al. (2012) [9] had implemented Support Vector Machines for finding the experts in the CQA system. They took stack overflow as dataset and had evaluated the system using Precision (P), Recall (R), F-Score (F). Tom Chao Zhou et al. (2012) [4] had implemented Support Vector Machine for bridging the gap between posted questions and potential answerers. They had used Yahoo Answering dataset and had demonstrated a high feasibility of question routing towards the expert over the parameters like P, R, F1, Accurcay (A). Aditya Pal et al. (2012) [10] had proposed the use of Probabilistic models for solving the issue of expert finding and to obtain the users that have the capability of becoming an expert in future using P, R, F as a performance parameters. Akshi Kumar et al. (2012) [11] had proposed a ComEx Miner System for ranking the blogs and then mining the experts based on the blog ranks. Guangyou Zhou et al. (2012) [12] had proposed a Topic-Sensitive Probabilistic model in which they had considered link structure technique while the other existing approaches had used link analysis techniques and showed that there technique performed well for finding the expert in CQA using Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Average Precision (AP) as performance parameters. Muhammad Asaduzzaman et al. (2013) [13] had implemented Random Forest and J48 for determining the duration till the question remains unanswered using P, R as performance parameters. Alessandro Bozzon et al (2013) [14] had implemented Vector Space model for determining the expert users in the CQA over the social network datasets using MAP, Point Interpolated Average Precision (PIAP), MRR, Discount Cumulative Gain (DCG) as performance parameters. Shuo Chang et al (2013) [15] had proposed an Expert Recommendation model in which they had focused on routing a question to a group of experts rather that routing it to a particular expert using P, R, F parameters for the random data sets. Chih-Lu Lin et al (2014) [16] had presented an approach named as Knowledge-gap-based difficulty rank algorithm for calculating the knowledge gap that existed in different CQA categories using F, area under the ROC curve (AUC), MAP, normalized discount cumulative gain (nDCG), precision for top 5 user (P@5) and Concordance Analysis (CA). The results showed that there approach performed better than the other existing approaches. Baoguo Yang et al (2014) [17] had proposed a new probabilistic model called User topical ability model for finding expert users based on the explanation of the answer and their descriptive ability based on question. For the experiment purpose, they took stack overflow as dataset and had calculated Mean Average Precision (MAP), Average Rank (AR). MRR, Cumulative Distribution of Ranks (CDR) for evaluating the performance. Baoguo Yang et al. (2014) [18] had proposed a model called Tag-Based Expert Recommendation model where they had used Question for finding expert users and showed that there method performed better than the earlier methods. For experimentation, they had used stack overflow and had calculated nDCG, P, Kendall (K) performance parameters for evaluation. Xiaoqiang Zhou et al. (2015) [19] had used Convolution Neural Network model for learning the question and answer pairs and had then implemented Long Short Term Memory (LSTM) for obtaining the quality of each answer. They took SemEval dataset. Majid Rafiei et al. (2015) [20] had proposed a method called Hybrid method for finding experts in online communities based on social network and content analysis and had evaluated the system using Spearman Corelation (SpCo). Zhou Zhao et al. (2015) [21] worked on the graph regularized matrix model for finding experts in CQA. They had further described two procedures (i) Graph regularized matrix completion expanded gradient method (GRMC-EGM) and (ii) Graph regularized matrix completion advanced proximal gradient search method (GRMC-AGM) for finding expert in Question Answering sites and had evaluated the system using MRR, DCG, Percision for Top 1 user (P@1) and A. They had used stack exchange dataset. Another author Ivan Srba et al. (2015) [22] discussed about the Question Routing Method using Non QA data for analysing users in order to estimate user interests and expertise in the early stages. They had used stack overflow dataset and had calculated MRR, Precision for Top 5 user (P@5), Precision for Top 10 user (P@10) performance parameters. Hai Li et al. (2016) [23] had proposed a Novel Hybrid Model that helped to find experts in the CQA based on considering user post, answer votes, best answer ratio and user relation. They had used the efficacy criteria's like Precision (P), Recall (R) and Reciprocal Rank (RR). Experimental results showed that the performance improved by 2.97% to 7.79%. Deba P. Mandal et al (2015) [24] had observed that there are large number of questions that are unanswered. So they had proposed a query likelihood language (QLL) model for solving the problem by routing a novel question to the expert users that have the potential to answer the questions. MRR, Best Answer Coverage (BAC), Success (S) are the parameters they had calculated for evaluation purposes and results were quite encouraging. Geerthik.S et al (2016) [25] had proposed a model called Domain Expert Ranking model that identified the expert in the CQA and results showed that it performed better than the existing expert finding model when measured on multiple metrics. They had used MRR as a performance parameter. Juan Yang et al. (2016) [26] had proposed a NEWHITS model and had compared it with traditional link analysis model. For evaluation of their system, they had used Mean Reciprocal Rank (MRR) and discount cumulative gain (DCG) as key performance indicator's (KPI's). Muhammad Ahasanuzzaman et al. (2016) [27] had proposed a method called Dupe that helped to find the duplicate questions when applied to the stack overflow dataset. They had compared the results with the DupePredictor model and showed that their model had performed better in terms of recall-rate (RR). Haifa Alharthi et al. (2016) [28] had also proposed a method called linear prediction model for predicting the score of the questions. These scores were based on sixteen factors. They had used stack overflow dataset and had evaluated the performance of the new method using Average Score (AS) and Median Score (MS). Jinwei Liu et al. (2017) [29] had proposed a Coupled Semi-Supervised Mutual Reinforcement based Label Propagation (CSMRLP) model that helped in improving the Quality of Service. Mohammad Javad Kargar et al (2017) [30] had proposed an open model based on crowdsourcing for a QA system.

Following table 1 illustrates the work done by various authors in the field of expert mining in CQA.

Table 2.1: Summary of Studies Related to Expert Mining in CQA

Author	**Year**	**Publication**	**Technique**	**Dataset**	**Performance parameter**
Javad Kargar M.	2017	ICWR	An open model	Yahoo	-

m Oveissi A. [30]			based on Crowd sourcing		
Liu J. , Shen H. [29]	2017	IEEE	Coupled Semi-Supervised Mutual Reinforcement-based Label Propagation (CSMRLP)	Yahoo answering	P, R, F1, A.
Alharthi H. , Outioua D. , Baysal O. [28]	2016	ACM	Linear prediction model.	Stack overflow	AS,MS
Ahasanuzzaman M. , K.Roy C. [27]	2016	MSR	Dupe	Stack overflow	RR
Yang J. , Peng S. , Wang L. , Wu B. [26]	2016	IEEE	NEWHITS algorithm	Stack overflow	MRR, DCG
Geerthis S. , Dr. K R Gandhi. , Venkatraman S [25]	2016	IEEE	Domain expert ranking	Quora	MRR
P. Mandal D. , Kundu D. , Maiti S. [24]	2015	ICACEA	Theme based Query Likelihood Language	Yahoo answering	MRR, BAC, S
Li H. , Jin S. , Li S. [23]	2015	IEEE	A novel Hybrid analysis model	Stack overflow	P, R, RR
Srba I. , Grznar M., Bielikova M. [22]	2015	IEEE/ACM	Question Routing method	Stack Exchange	MRR, P@5, P@10
Zhxao Z. , Zhang L. He X. Ng W. [21]	2015	IEEE	Graph regularized matrix completion algorithm	Quora	MRR, DCG, P@1,

					A.
A Kardan A. , Rafei M. [20]	2015	Springer	A Hybrid method	Java Online Communities	SpCo
Zhou X., Hu B., Chen Q, Tang B., Wang X. [19]	2015	ACL-IJCNLP	Convolution Neural Network	SemEval	P, R, F1
Yang B. , Manandhar S. [18]	2014	IEEE/ACM	Tag-Based Expert Recommendation Model	Stack Overflow	nDCG, P, K
Yang B. , Manandhar S. [17]	2014	IEEE/ACM	User topical ability model	Stack Overflow	MAP, AR, MRR, CDR.
Lin C.Lu, Chen Y-Liang, and Kao K-Yu [16]	2014	IEEE/ACM	knowledge-gap-based difficulty rank (KG-DRank) algorithm	Yahoo answering	F, AUC, MAP, nDCG, P@5 ,CA
Pal A., Chang S. [15]	2013	IEEE/ACM	Expert Recommendation model	Random Data Set	P, R, F
Ren Liu D., Hsuan Chen Y., Chen Kao W., Wen Wang H. [31]	2013	Information processing and management	Vector space model	Yahoo answering	MRR, P@5, MAP
Bozzon A., Brambilla M. , Ceri S., Silvestri M. , Vesci G. [14]	2013	ACM	Vector space model	Social network	MAP, PIAP, MRR, DCG.
Asaduzzaman M., Shah Mashiyat A., K. Roy C. , A. Schneider K. [13]	2013	IEEE	Random Forest, J48	Stack Overflow	P, R.
Zhou G. , Lai S. , Liu K. , ftZhao J.	2012	ACM	Topic-sensitive probabilistic	Yahoo answering	MAP, MRR, AP.

[12]			model		
Kumar A., Ahmad N. [11]	2012	IJACSA	ComEx Miner System	Blogs	A
PAL A., HARPER M. F., A. KONSTAN J. [10]	2012	ACM	Probabilistic model	Turbo tax live community, stack overflow	P, R, F.
Chao Zhou T. , R. Lyu1 M., King1 I. [4]	2012	ACM	Support vector machine	Yahoo answering	P, R, F1, A.
Pal A., Chang S., A. Konstan J. [9]	2012	AAAI	Gaussian Mixture Model	Stack overflow	P, R, F.
Pal A., A. Konstan J.	2010	ACM	Machine Learning Models	Turbo tax live communities, Stack Overflow	P, R, F1.
Shah C., Pomerantz J. [7]	2010	ACM	Logistic Regression Model	Yahoo Answering	M, SE
Lu y. , Quan X. , Ni X. , Liu W. , Xu Y. [6]	2009	IEEE	Latent link analysis approach	Yahoo answering	P
Bouguessa M., Dumoulin B. , Wang S [5]	2008	ACM	Probabilistic model	Yahoo answering	-

2.2 Perspective of Expert Mining in CQA

An online community or a virtual community or an internet community is often been described as a place where people interact with each other using Internet [11]. In CQA community people can post, comment on discussions, collaborate or give advices. The most common among them are the social networking [32] sites like Twitter etc., E-mails, forums, chat rooms, blogs, video games etc. through which people can actively communicate with each other [33] . There are three categories of virtual communities [34]. They are social orientation, professional orientation and commercial orientation. Social oriented virtual communities are related to the social network such as relationship building, entertainment etc. Professional orientation are those virtual communities that deals with professional environment such as expert network etc. Commercial orientation virtual communities are related to the information provided by the virtual communities such as Trade, geographical proximity, brand / product support etc.

In order to perceive, interpret and understand principles, theories and concepts of CQA systems, it is recognized that the question-answering process can be comprehended from two different axes [35] i.e. the first perspective considers CQA as an information sharing that is basically based on knowledge sharing. Knowledge sharing is referred as a process in which knowledge is exchanged among members of a particular community. Second perspective considers CQA as a searching system that aims to search a question based on a very specific way of informal learning.

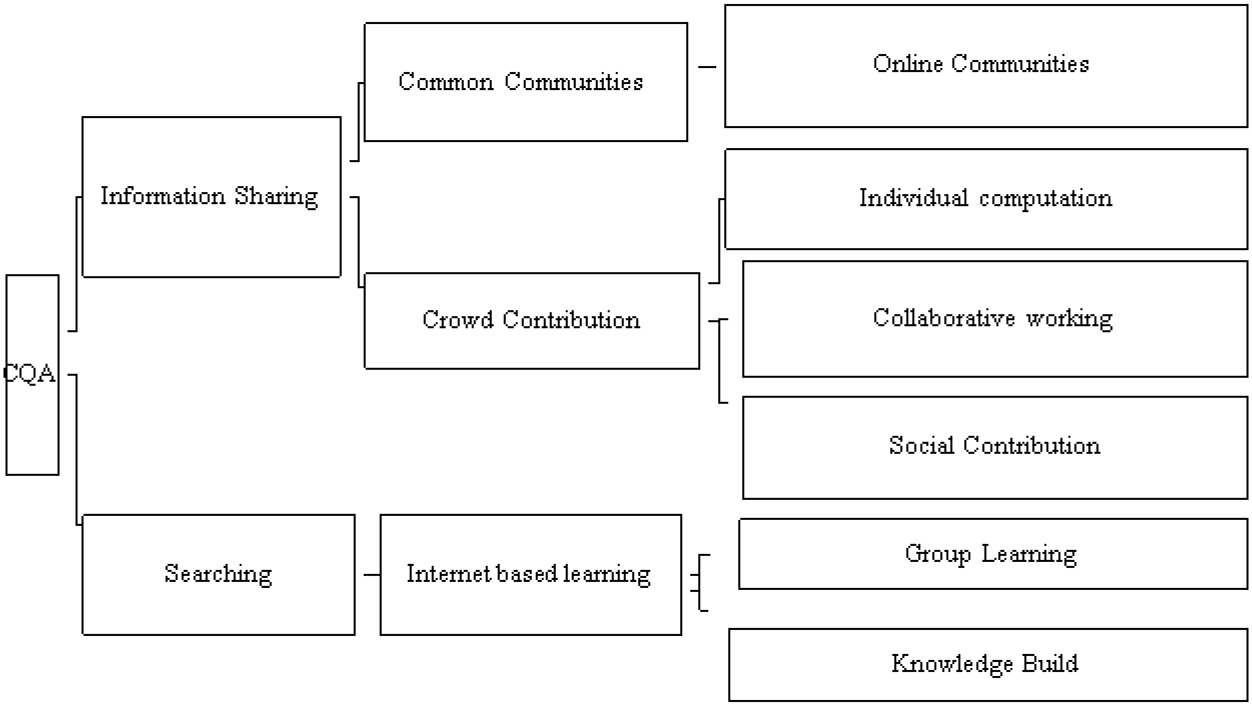


Fig. 2.1 Perspectives of CQA

Figure 2.1 illustrates the two perspectives of CQA. The first perspective i.e. “Information Sharing” or the knowledge sharing [35] can be further divided into two parts as follows: (i) people who have common interests in a particular subject is known as “Common Communities”.

As all the users interact with each other using Internet, we call this community as an Online Community. (ii) “Crowd Contribution” is based on a simple idea that nobody knows everything. This motivates the whole community to look forward and harness for ideas together for solving the problems that would be quite difficult to solve individually. Crowd Contribution represent three subsequent theories that are related to CQA systems: “Individual Computation, Collaborative Working, and Social Contribution”. The second perspective is “Searching” [35]. People can explore, search and learn things from the experts by exploring, searching, reading, asking and answering questions. Usually in the CQA there is no teacher or instructor. This type of learning is called Internet based learning where learners are sharing their knowledge with each other instead of getting any guidance from teacher or instructor over the Internet. Groups of people can share their knowledge worldwide using Internet.

2.3 Approaches towards Expert Finding in CQA

There are two approaches that can be used for expert finding in CQA. The first is “Graph Based” approach and the second is “Featured Based” approach.

The Graph Based is represented by an expertise graph where nodes represent the “Domain” entities and the edges between the nodes denotes the “Notion of Expertise”. Here “Domain” entities denotes whether the user is expert or non-expert and the “Notion of Expertise” denotes the connection, prominence, trust, etc. The Graph Based approach utilizes the application of the popular algorithms like PageRank, Link analysis, NEWHITS, Graph regularized matrix completion algorithm, Domain expert ranking etc. Figure 2.2 depicts the Graph Based approach.

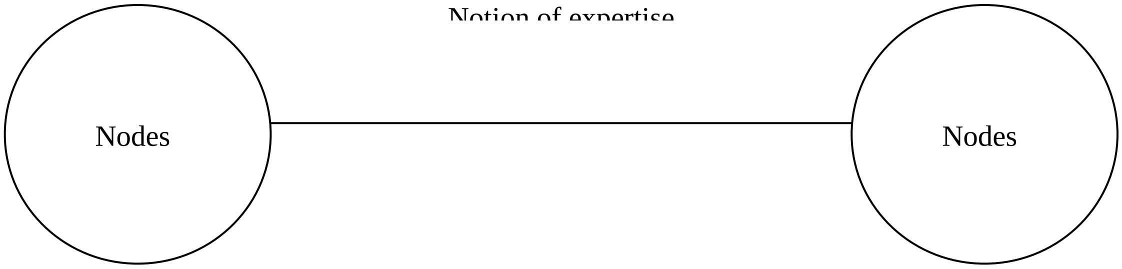


Fig. 2.2 Representation of Graph based approach

The most famous Graph Based approaches are the PageRank algorithm and HITS algorithm. Both the algorithms model the Internet as a graph. Node represents a Web page and the edges represents the hyperlink between two pages. This representation is also called Web graph. The PageRank algorithm is based on probability of a random web page. This algorithm is based on Markov chains. On the other hand, HITS algorithm computes two values namely authority value and hub value for a Web page. The authority value is the value of content of the page and the hub value is the value of a page links to other pages. Several other algorithms also has been proposed based on these two algorithms with slight variants such as SALSA, EntityRank, TwitterRank and AuthorRank.

The Feature Based approaches select features for each entity based on particular domain i.e. the main features that can be extracted for finding experts in CQA are the number of answers, the number of questions, part-of-speech (POS) analysis, the number of reputations, the number of up votes and down votes, graph features, etc. The addition features are also computed using these features like Z-score. Then these features are used for identifying experts in the CQA using machine learning algorithms like clustering algorithms, ranking algorithms and/or using thresholds based on domain knowledge. The featured based approach mentions the application of algorithms like Vector space model, Probabilistic model, Gaussian mixture model, Random Forest, Linear prediction model, J48 etc. Figure 2.3 represents the Feature Based approaches.

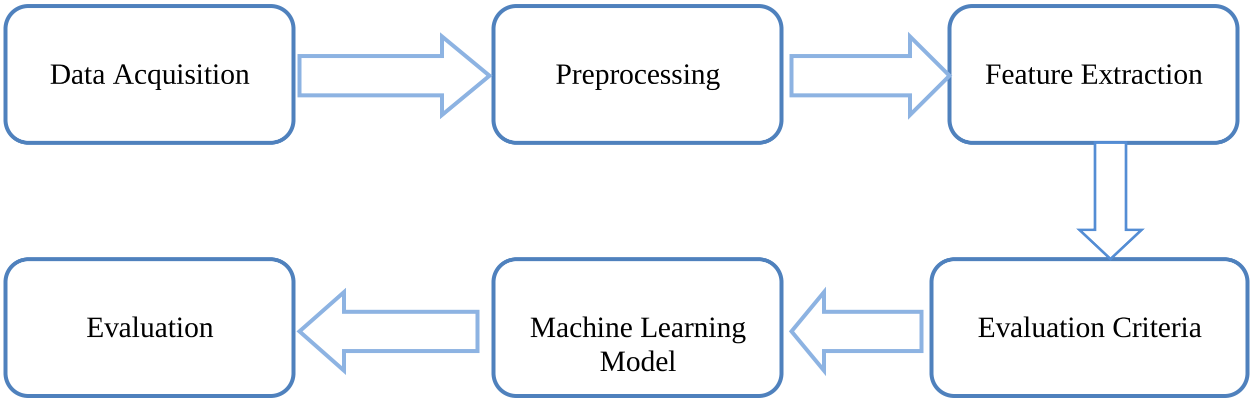


Fig. 2.3 Representation of Featured Based approach

Figure 2.4 depicts the various methods that can be used for expert finding. The Graph Based approach utilizes the application of the popular algorithms like PageRank, Link analysis, NEWHITS, Graph regularized matrix completion algorithm, Domain expert ranking etc. whereas the featured based approach mentions the application of algorithms like Vector space model, Probabilistic model, Gaussian mixture model, Random Forest, Linear prediction model, J48 etc.

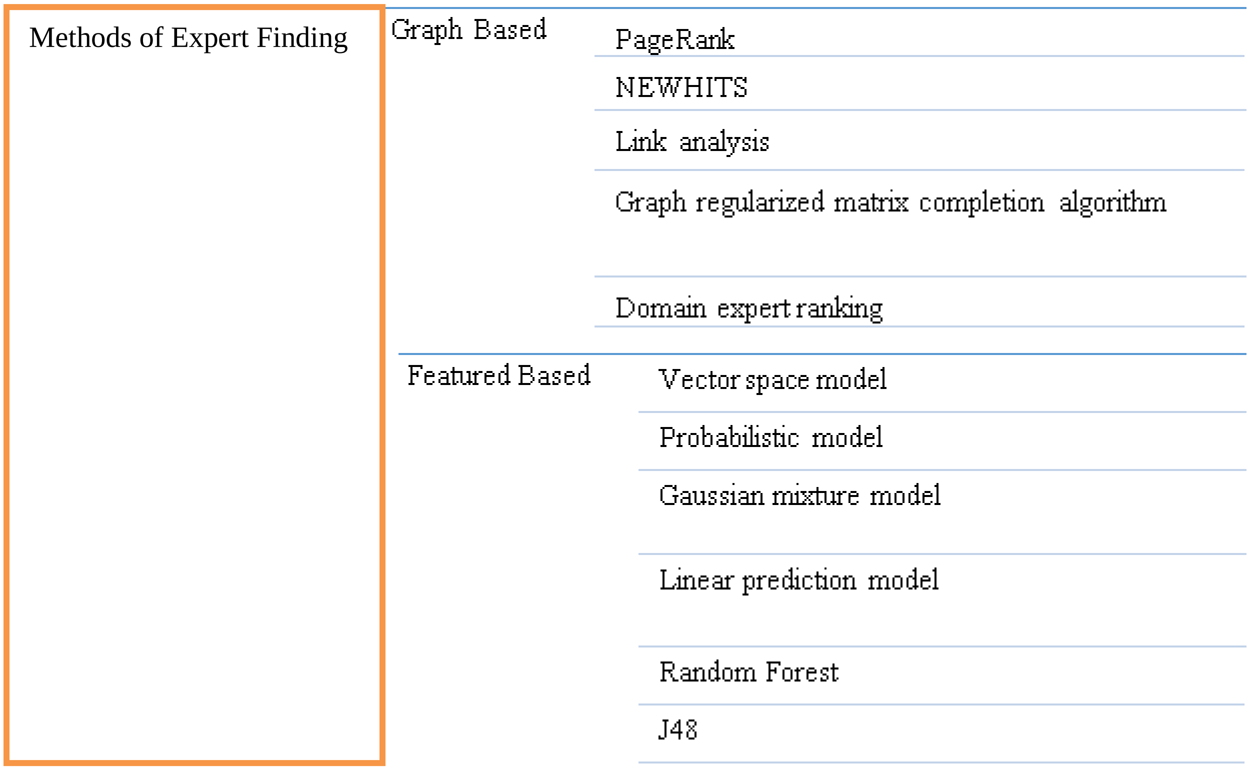


Fig. 2.4 Various methods used for Expert Finding

Chapter 3 Proposed Model

Chapter 2 identified issues related to the expert finding in CQA systems. This chapter illustrates the novel techniques that constitute the proposed model to address those issues presented in Chapter 2. Section 3.1 illustrates the proposed model, describes each component of the system and shows how each of the proposed technique contributes to the expert finding in the CQA.

3.1. The Proposed Model

The proposed model has primarily five phases

i. Data Acquisition
ii. Pre-processing and Feature Extraction
iii. Evaluation Criteria
iv. Learning Model
v. Evaluation

In the initial step, the data set is collected from Stack Overflow. It includes over millions of data. To limit the scope, 5000 data items were used. The pre-processing task includes cleaning, transformation and reduction. The relevant features (including id, reputation, up-vote, down-vote, creation-date, last-access-date) are then extracted and then the tenure for each user is calculated. After that, each user is either classified as expert or non-expert using the formula "((up_vote – down_vote) + (reputation/tenure))" and then various supervised learning techniques namely, NB (Naïve Bayes), SVM (Support Vector Machine), RF (Random Forest), K-NN (K-Nearest Neighbor) and K-STAR are applied on the dataset. In the final step, results obtained are then evaluated using various performance measures.

The following figure 3.1 illustrates the model:

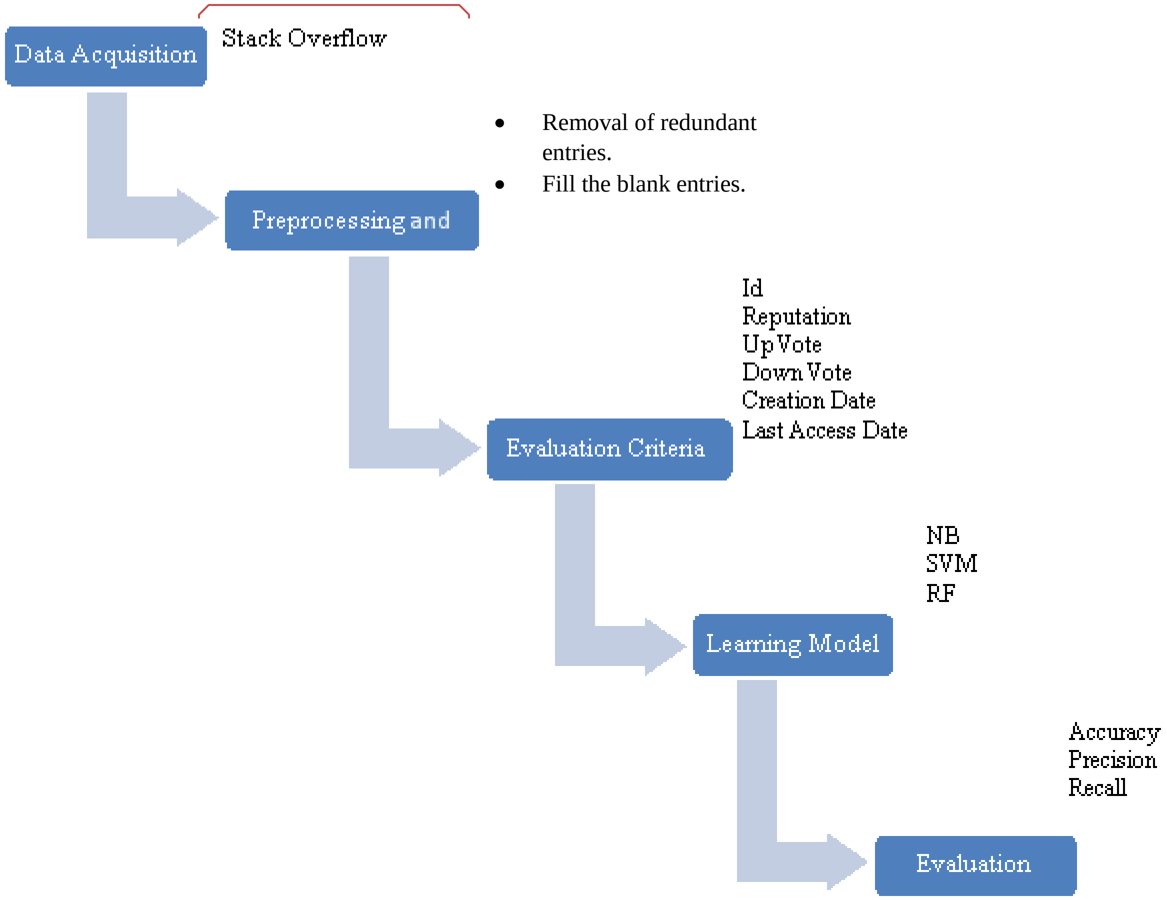


Fig. 3.1 The predictive learning model for the CQA system

3.1.1 Dataset

The dataset used for the classification of users as experts and non-experts is formed by taking publicly available user's profile data from August 2008. The attributes considered in the dataset are Id, Reputation, Up Vote, Down Vote, Creation Date and Last Access Date. Id represents the unique id of each user. Reputation is the extent to which the user is trusted by the community of Stack Overflow. A question get an "Up Vote" if user considers it as a good question and an answer get an "Up Vote" if the answer is relevant and accurate to the asked query. Similarly, a question or an answer gets a "Down Vote" if user finds them irrelevant or inaccurate. "Creation Date" implies the time the profile is created and "Last Access Date" conveys the time the user profile is last accessed by him.

The dataset used for this study contains 5000 data items. 70% of the data items i.e. 3500 data items were used for training and rest 30% of the data items i.e. 1500 data items were used in the testing phase.

The following table 3.1 represents the snippet of dataset used.

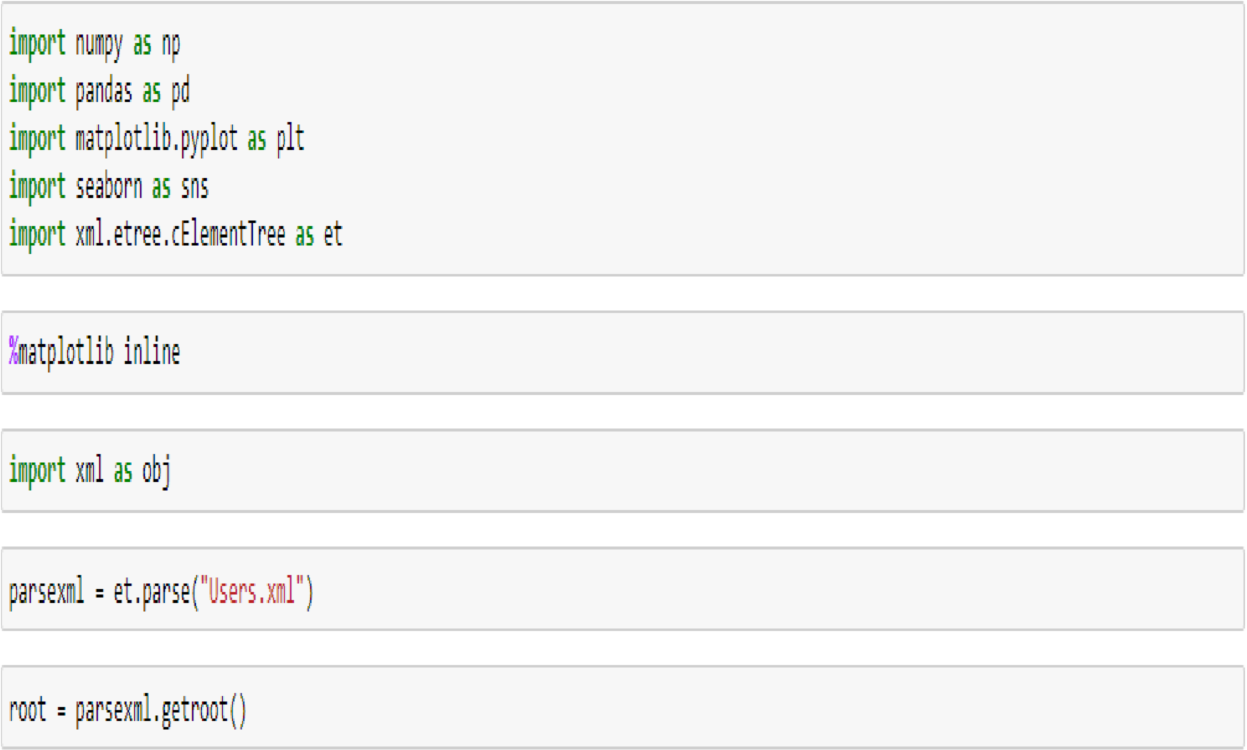
Table 3.1 Snippet of the dataset used

Id	Creation_date	Last_access_date	Up_vote	Down_vote	Reputation	tenure	Classification
5880845	2016-02-04 01:39:58.707000 UTC	2016-10-28 20:14:18.660000 UTC	0	0	15	812	N
2365460	2013-05-09 08:24:45.786999 UTC	2018-03-08 09:40:05.690000 UTC	2536	0	484	1813	Y
5821790	2016-01-21 14:39:36.869999 UTC	2018-03-08 16:00:50.769999 UTC	757	2	360	826	Y
970909	2011-09-29 10:22:46.592999 UTC	2018-03-09 07:29:28.423000 UTC	12	3	141	2401	N
1729210	2012-10-08 15:00:21.930000 UTC	2017-02-24 09:31:02.279999 UTC	16	3	281	2026	N
2691647	2013-08-17 09:08:13.197000 UTC	2018-03-04 21:44:30.973000 UTC	1738	6	390	1713	Y
176752	2009-09-21 19:28:47.822999 UTC	2018-03-08 13:25:59.877000 UTC	129	12	342	3138	N
442221	2010-09-08 08:43:46.130000 UTC	2018-03-09 15:28:36.763000 UTC	31	0	569	2787	N
12451	2008-09-16 14:25:04.933000 UTC	2018-03-06 19:19:34.009999 UTC	894	10	617	3509	Y
562692	2011-01-04 14:44:35.806999 UTC	2018-03-08 20:37:30.852999 UTC	913	2	1797	2669	Y

3.1.2 Preprocessing and Feature Extraction

The next step is to pre-process the acquired data. In this step the unwanted attributes like display name, website URL, about me, views, profile image URL, account id, location and age were removed as they were not required for this model. The redundant entries were removed and blank entries were assigned the minimum value for ease in assessment. The only required fields for the model are Id, Reputation, Up Vote, Down Vote, Creation Date and Last Access Date. The following figure represent the data before preprocessing.

Fig. 3.2 Dataset before preprocessing



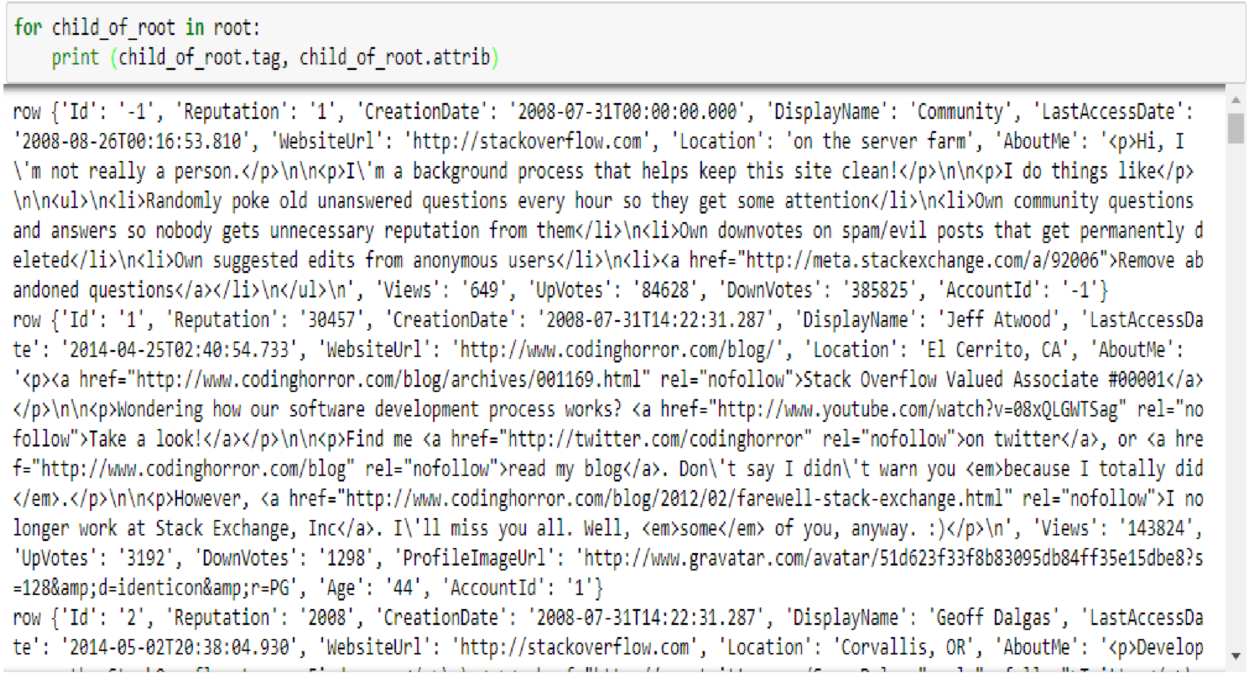
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import xml.etree.cElementTree as et
```

```
%matplotlib inline
```

```
import xml as obj
```

```
parsexml = et.parse("Users.xml")
```

```
root = parsexml.getroot()
```



```
for child_of_root in root:
    print (child_of_root.tag, child_of_root.attrib)
```

```
row {'Id': '-1', 'Reputation': '1', 'CreationDate': '2008-07-31T00:00:00.000', 'DisplayName': 'Community', 'LastAccessDate':
'2008-08-26T00:16:53.810', 'WebsiteUrl': 'http://stackoverflow.com', 'Location': 'on the server farm', 'AboutMe': '<p>Hi, I
\'m not really a person.</p>\n\n<p>I\'m a background process that helps keep this site clean!</p>\n\n<p>I do things like</p>
\n\n<ul>\n<li>Randomly poke old unanswered questions every hour so they get some attention</li>\n<li>Own community questions
and answers so nobody gets unnecessary reputation from them</li>\n<li>Own downvotes on spam/evil posts that get permanently d
eleted</li>\n<li>Own suggested edits from anonymous users</li>\n<li><a href="http://meta.stackexchange.com/a/92006">Remove ab
andoned questions</a></li>\n</ul>\n', 'Views': '649', 'UpVotes': '84628', 'DownVotes': '385825', 'AccountId': '-1'}
row {'Id': '1', 'Reputation': '30457', 'CreationDate': '2008-07-31T14:22:31.287', 'DisplayName': 'Jeff Atwood', 'LastAccessDa
te': '2014-04-25T02:40:54.733', 'WebsiteUrl': 'http://www.codinghorror.com/blog/', 'Location': 'El Cerrito, CA', 'AboutMe':
'<p><a href="http://www.codinghorror.com/blog/archives/001169.html" rel="nofollow">Stack Overflow Valued Associate #00001</a>
</p>\n\n<p>Wondering how our software development process works? <a href="http://www.youtube.com/watch?v=08xQLGWTSag" rel="no
follow">Take a look!</a></p>\n\n<p>Find me <a href="http://twitter.com/codinghorror" rel="nofollow">on twitter</a>, or <a hre
f="http://www.codinghorror.com/blog" rel="nofollow">read my blog</a>. Don\'t say I didn\'t warn you <em>because I totally did
</em>.</p>\n\n<p>However, <a href="http://www.codinghorror.com/blog/2012/02/farewell-stack-exchange.html" rel="nofollow">I no
longer work at Stack Exchange, Inc</a>. I\'ll miss you all. Well, <em>some</em> of you, anyway. :)</p>\n', 'Views': '143824',
'UpVotes': '3192', 'DownVotes': '1298', 'ProfileImageUrl': 'http://www.gravatar.com/avatar/51d623f33f8b83095db84ff35e15dbe8?s
=128&amp;d=identicon&amp;r=PG', 'Age': '44', 'AccountId': '1'}
row {'Id': '2', 'Reputation': '2008', 'CreationDate': '2008-07-31T14:22:31.287', 'DisplayName': 'Geoff Dalgas', 'LastAccessDa
te': '2014-05-02T20:38:04.930', 'WebsiteUrl': 'http://stackoverflow.com', 'Location': 'Corvallis, OR', 'AboutMe': '<p>Develop
```

3.1.3 Evaluation Criteria

In this step, the required fields were extracted to predict the expert or non-expert. Then the tenure of each user is calculated. Tenure is calculated using the formula that takes into consideration creation date and present date (March 2018). After that, the formula ((up_vote – down_vote) + (reputation/tenure)) is used to classify each user either as an expert or non-expert.

3.1.4 Learning Model

In this phase, the supervised learning techniques were used. The details about these techniques are given in the table 3.2.

Table 3.2: Supervised learning techniques used

Technique	**Description**
Naive Bayes	It is a very basic classifier model in machine learning which uses “probabilistic classifiers” based on applying Baye’s Theorem.
SVM	SVM or Support Vector Machines are supervised learning algorithms that are used to classify data by mapping the data

	points in n- dimensional space, separating them by finding the hyper-plane that segregates the classes and predicting the categories based on the class it falls in.
RF	Random Forest is an ensemble learning method. At the time of training, it uses bagging method i.e. it creates a number of decision trees employing different learning methods and find the result by voting for the most popular class. The two popular methods of Random Forest are Boosting and Bagging of the classification trees.
K-NN	K-NN (K-Nearest Neighbor) is a non-parametric supervised learning algorithm that stores all available learning data points in the feature space and classifies new data points based on a similarity measure like distance function. A data item is classified into a class by a majority vote of its neighbors with the data item being allotted to the class most frequent among its k nearest neighbors.
K-STAR	K-Star algorithm uses entropy for classification. This algorithm is based on probability. It transform instance into another by randomly choosing between all possible transformations.

3.1.2 Evaluation

The results are observed and analyzed using few parameters such as precision(P), recall(R), accuracy(A) and F-measure(F). The results are discussed in the next section.

This sections highlights the efficacy measures used in the experiments like P, R, A and F.

Table 3.3: Efficacy Measures used

Measure	Description
Accuracy	Accuracy refers to the closeness of a measured or predicted value to a standard value. It is the ratio of the correctly

	predicted values to the total number of predictions.
Precision	Precision (also known as Positive Predictive Value or Confidence) is the proportion of the correctly predicted positive values to the predicted positive values.
Recall	Recall (or Sensitivity) is the proportion of the correctly predicted positive values to the actual positive values. It assists in computing coverage of real positive cases.
F- Measure	The F measure is described as the weighted mean of the precision and recall of the data. It's a combined metric which determines the effectiveness of a data under observation. It is calculated as the ratio of twice the product of precision and recall to the sum of precision and recall.

Chapter 4 Experimental Results and Analysis

This chapter describes the experimental results and the analysis to account for the tests performed.

Table 4.1 Performance measure results for different ML techniques

MESURES TECHNIQUES	A	P	R	F
NB	84.09	.83	.84	.83
SVM	83.28	.83	.83	.84
RF	86.23	.86	.87	.86
K-NN	83.79	.82	.83	.83
K-Star	76	.75	.74	.75

From the above table 4.1, it is observed that Random Forest and Naive Bayesian gives the highest accuracy score (86.23% and 84.09% respectively). After these, the K-NN model showed 83.79% accuracy and SVM showed 83.28% accuracy. K-Star showed the lowest accuracy of around 76%.

The following figure 4.1, 4.2, 4.3 and 4.4 represents the results with the help of graphs.

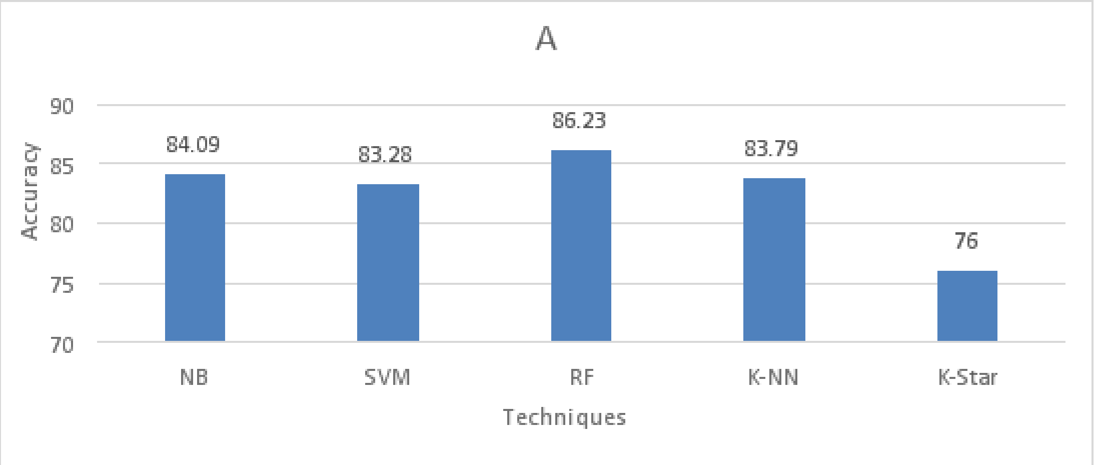


Fig. 4.1 Accuracy

Accuracy of various techniques is presented in figure 4.1. From this, it is concluded that Random Forest and Naïve Bayes gives the highest accuracy score i.e. 86.23 and 84.09 respectively.

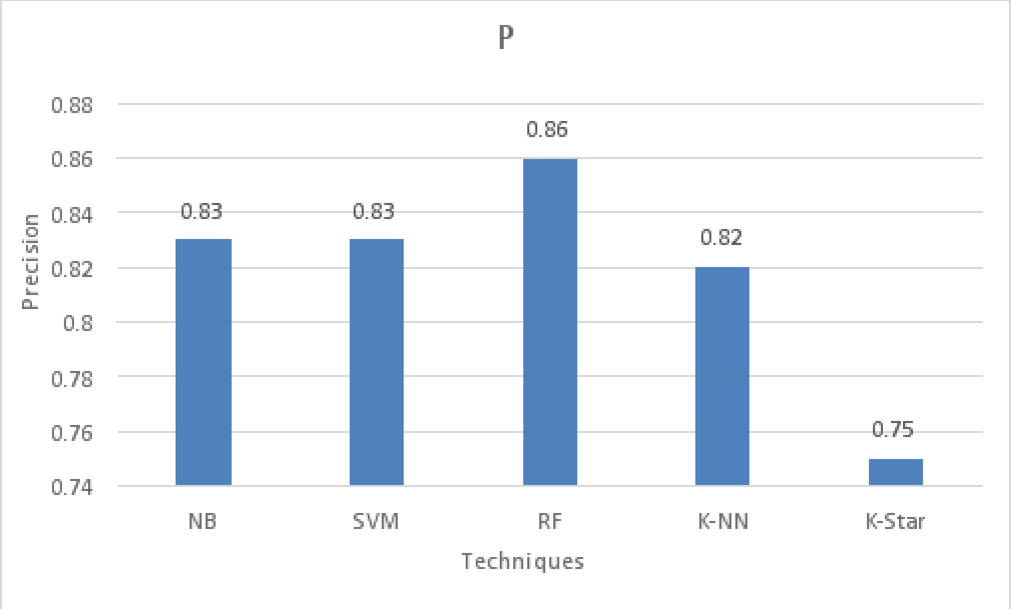


Fig. 4.2 Precision

Precision of various techniques is presented in figure 4.2. From this, it is concluded that Random Forest yield the highest precision score i.e. 86. After that Naïve Bayes and SVM have the same precision i.e. .83 and .83 respectively.

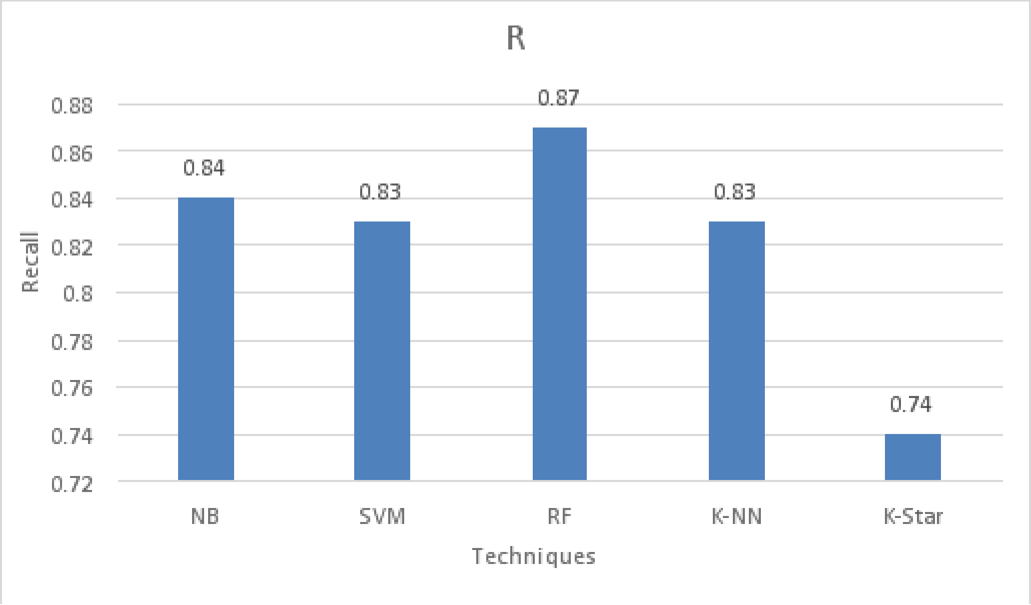


Fig. 4.3 Recall

Recall of various techniques is presented in figure 4.3. From this, it is concluded that Random Forest and Naïve Bayes yields the highest recall score i.e. .87 and .84 respectively.

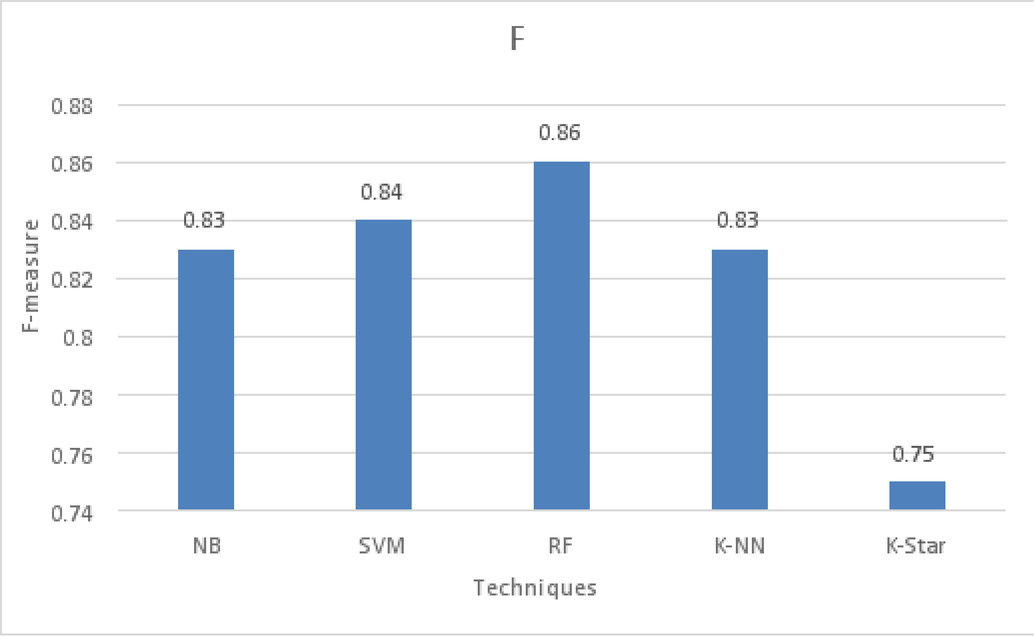


Fig. 4.4 F-measure

F-measure of various techniques is presented in figure 4.4. From this, it is concluded that Random Forest gives the highest F-measure i.e. .86 and K-Star gives the lowest F-measure i.e. .75.

Chapter 5 Conclusion

In this project, we produced competitive results on the Stack Overflow dataset using five machine learning algorithms namely Naïve Bayes, Support Vector Machine, Random Forest, k-Nearest Neighbor and K-Star have encouraging results, with Random Forest and Naïve Bayes having the highest accuracy with 86.23 and 84.04 respectively. For Recall as the performance measure, Random Forest and Naïve Bayes achieved the highest Recall values with .87 and .84 respectively. Among all, K-Star gives the lowest results lagged behind the other four algorithms. Overall, we can deduce to the conclusion that the results are quite appreciating with Random Forest and Naïve Bayes yielding the best results for this model.

As it can be concluded that finding of experts in Stack Overflow using machine learning techniques yield appreciable results. Thus this model can be applied on other question answering systems like Quora, Yahoo! Answering. Also, attributes like tags and badges could be considered to evaluate experts with greater accuracy.

Bibliography

[1] N. Pathak, B. M. Singh, and G. Sharma, “UML 2.0 based framework for the development of secure web application,” *Int. J. Inf. Technol.*, vol. 9, no. 1, pp. 101–109, Mar. 2017.

[2] N. Narwal, S. K. Sharma, and A. P. Singh, “Web content adaptation using 2D bin packing algorithm,” *Int. J. Inf. Technol.*, vol. 9, no. 2, pp. 139–146, Jun. 2017.

[3] J. Manhas, “Initial framework for website design and development,” *Int. J. Inf. Technol.*, vol. 9, no. 4, pp. 363–375, Dec. 2017.

[4] T. C. Zhou, M. R. Lyu, and I. King, “A classification-based approach to question routing in community question answering,” in *Proceedings of the 21st international conference companion on World Wide Web - WWW ’12 Companion*, 2012, p. 783.

[5] M. Bouguessa, B. Dumoulin, and S. Wang, “Identifying authoritative actors in question-answering forums,” in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 2008, p. 866.

[6] Y. Lu, X. Quan, X. Ni, W. Liu, and Y. Xu, “Latent link analysis for expert finding in user-interactive question answering services,” in *SKG 2009 - 5th International Conference on Semantics, Knowledge, and Grid*, 2009, pp. 54–59.

[7] C. Shah and J. Pomerantz, “Evaluating and predicting answer quality in community QA,” in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’10*, 2010, p. 411.

[8] A. Pal, “Expert identification in community question answering: exploring question selection bias,” *Proc. 19th ACM Int.*, pp. 1505–1508, 2010.

[9] A. Pal, S. Chang, and J. A. Konstan, “Evolution of Experts in Question Answering Communities,” *Proc. Sixth Int. AAAI Conf. Weblogs Soc. Media*, pp. 1–8, 2012.

[10] A. Pal, F. M. Harper, and J. A. Konstan, “Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering,” *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 1–28, 2012.

[11] A. Kumar and N. Ahmad, “ComEx Miner: Expert Mining in Virtual Communities,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 6, 2012.

[12] G. Zhou, S. Lai, K. Liu, and J. Zhao, “Topic-sensitive probabilistic model for expert finding in question answer communities,” in *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM ’12*, 2012, p. 1662.

[13] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider, “Answering questions about unanswered questions of Stack Overflow,” *2013 10th Work. Conf. Min. Softw. Repos.*, pp. 97–100, 2013.

[14] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, “Choosing the Right Crowd : Expert Finding in Social Networks Categories and Subject Descriptors,” *Proc. 16th Int. Conf. Extending Database Technol.*, pp. 637–348, 2013.

[15] S. Chang and A. Pal, “Routing questions for collaborative answering in community question answering,” *Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. - ASONAM ’13*, pp. 494–501, 2013.

[16] C. L. Lin, Y. L. Chen, and H. Y. Kao, “Question difficulty evaluation by knowledge gap analysis in Question Answer communities,” in *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014, pp. 336–339.

[17] B. Yang and S. Manandhar, “Exploring user expertise and descriptive ability in community question answering,” in *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining, 2014, pp. 320–327.

[18] B. Yang and S. Manandhar, “Tag-based expert recommendation in community question answering,” in *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014, pp. 960–963.

[19] X. Zhou, B. Hu, Q. Chen, B. Tang, and X. Wang, “Answer Sequence Learning with Neural Networks for Answer Selection in Community Question Answering,” *Acl-2015*, pp. 713–718, 2015.

[20] M. Rafiei and A. A. Kardan, “A novel method for expert finding in online communities based on concept map and PageRank,” *Human-centric Comput. Inf. Sci.*, vol. 5, no. 1, p. 10, 2015.

[21] Z. Zhao, L. Zhang, X. He, and W. Ng, “Expert Finding for Question Answering via Graph Regularized Matrix Completion,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 993–1004, 2015.

[22] I. Srba, M. Grznar, and M. Bielikova, “Utilizing Non-QA Data to Improve Questions Routing for Users with Low QA Activity in CQA,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM ’15*, 2015, pp. 129–136.

[23] H. Li, S. Jin, and S. Li, “A Hybrid Model for Experts Finding in Community Question Answering,” in *Proceedings - 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015*, 2015, pp. 176–184.

[24] D. P. Mandal, D. Kundu, and S. Maiti, “Finding experts in community question answering services: A theme based query likelihood language approach,” in *Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015*, 2015, pp. 423–427.

[25] S. Geerthik, K. Rajiv Gandhi, and S. Venkatraman, “Domain expert ranking for finding domain authoritative users on community question answering sites,” in *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, 2017.

[26] J. Yang, S. Peng, L. Wang, and B. Wu, “Finding experts in community question answering based on topic-sensitive link analysis,” in *Proceedings - 2016 IEEE 1st International Conference on Data Science in Cyberspace, DSC 2016*, 2017, pp. 54–60.

[27] M. Ahasanuzzaman, M. Asaduzzaman, C. K. Roy, and K. A. Schneider, “Mining duplicate questions in stack overflow,” in *Proceedings of the 13th International Workshop on Mining Software Repositories - MSR ’16*, 2016, pp. 402–412.

[28] H. Alharthi, D. Outioua, and O. Baysal, “Predicting questions’ scores on stack overflow,” in *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering - CSI-SE ’16*, 2016, pp. 1–7.

[29] J. Liu, H. Shen, and L. Yu, “Question quality analysis and prediction in community question answering services with coupled mutual reinforcement,” *IEEE Trans. Serv. Comput.*, vol. 10, no. 2, pp. 286–301, 2017.

[30] M. J. Kargar and A. Oveissi, “An open model for question answering systems based on Crowdsourcing,” in *2017 3rd International Conference on Web Research, ICWR 2017*, 2017, pp. 122–127.

[31] D. R. Liu, Y. H. Chen, W. C. Kao, and H. W. Wang, “Integrating expert profile, reputation and link analysis for expert finding in question-answering websites,” *Inf.*

Process. Manag., vol. 49, no. 1, pp. 312–329, 2013.

[32] I. Khan, S. K. Naqvi, M. Alam, and S. N. A. Rizvi, “An efficient framework for real-time tweet classification,” *Int. J. Inf. Technol.*, vol. 9, no. 2, pp. 215–221, Jun. 2017.

[33] T. K. Tran and T. T. Phan, “Mining opinion targets and opinion words from online reviews,” *Int. J. Inf. Technol.*, vol. 9, no. 3, pp. 239–249, Sep. 2017.

[34] C. E. Porter, “A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research,” *J. Comput. Commun.*, vol. 10, no. 1, pp. 00–00, 2006.

[35] I. Srba and M. Bielikova, “A Comprehensive Survey and Classification of Approaches for Community Question Answering,” *ACM Trans. Web*, vol. 10, no. 3, pp. 1–63, 2016.

Appendix A

List of Publications

Conference

1. A. Kumar and D. K. Gupta*, "Expert Mining In Community Question Answering*,"INDIACom-2018, IEEE Conference ID:4285 (Accepted, to be pubished).

2. A. Kumar and D. K. Gupta*, "Empirical Analysis of Supervised ML Techniques For Expert Mining In CQA Systems*"(Communicated).