IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 7, 2016, accepted August 6, 2016, date of publication August 26, 2016, date of current version September 16, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2600622

Selecting Best Answer: An Empirical Analysis on Community Question Answering Sites

TIRATH PRASAD SAHU[1], NARESH KUMAR NAGWANI[2], AND SHRISH VERMA[3]
[1]Department of Information Technology, National Institute of Technology Raipur, Raipur 492010, India
[2]Department of Computer Science and Engineering, National Institute of Technology Raipur, Raipur 492010, India
[3]Department of Electronics and Telecommunication Engineering, National Institute of Technology Raipur, Raipur 492010, India

Corresponding author: T. P. Sahu (tirsahu.it@nitrr.ac.in)

ABSTRACT A community question answering (CQA) site is a well-known online community, where user interacts on a wide variety of topics. To the best of our knowledge, the selection of a best answer for the question asked on the CQA site is done manually, which is traditional and tedious. In this paper, a model is developed for selecting best answer for the question asked on the CQA site. Instead of taking data related to question–answer only into account as done in manual process, this model takes both question–answer and answerers' data into account, which gives an insight view into the answers given by the experts that is more likely to be selected as the best answer. The presented approach analyzes StackOverflow Q&A posts with at least five answers to extract features for pattern identification using which the best answer is selected for the asked questions based on topic modeling and classifier. To evaluate correctness of the proposed model, a set of parameters are used, such as Receiver Operating Characteristics Area Under Curve, Precision Recall Area Under Curve, Gmean, and Accuracy. Results show that the proposed model is effective in predicting the best answer.

INDEX TERMS Classifier, community question answering (CQA), feature identification, online community, statistical analysis, topic modelling.

I. INTRODUCTION

Community question answering (CQA) sites have become an important source of content creation over the years, as these CQA sites have exceeded the rate of content consumption. Users ranging from naive to outshiners, visit CQAs to gain knowledge and seek answers to various type of questions [1]. There has been a rapid increase in two types of CQA (i) general question-answer sites (such as Quora[1] & Yahoo! Answers)[2] (ii) domain specific question-answer sites (such as StackOverflow[3] and AskUbuntu)[4] which have catered to programming related questions and turned into repositories of software engineering knowledge. StackOverflow is an interactive CQA site for software knowledge by hosting collaborative network of millions of users (developers). The users can create free account to: ask/answer questions, upvote/downvote questions/answers, gain reputation etc. to become an expert in the domain of software engineering.

In today's age, every day users are engaged on CQA sites with myriads of questions and their corresponding answers, which in turn, lead to continuous growing size of contents on this site. Such growing contents are posing several challenges which open up new opportunities for researchers to comprehend and establish meaningful patterns from the large size of contents that are available on such CQA sites. Analyzing and understanding the StackOverflow knowledge repository could provide key insights about domain knowledge (topical interest), activeness, expertise etc. that will help developers to choose the questions being answered in their working environments. However, it is tedious to analyze such huge and semi-structured textual contents along with associated post scores manually [2].

There are two types of users involved on StackOverflow (i) Asker – a user who posts question on a wide array of topics and wait an answer from other users. (ii) Answerer – a user who posts answer to the question posted by the asker. StackOverflow allows askers to pose their queries as questions and receives multiple answers from their fellow users. Posts on StackOverflow suggest that the first answer on a question arrives in about 9 minutes. The best answer to be accepted is the answer that satisfies the asker's question, usually arrives in a span of minutes. In this work, we study the answers received for the question and their related metrics for pattern

[1] https://www.quora.com/
[2] https://answers.yahoo.com/
[3] http://stackoverflow.com/
[4] http://askubuntu.com/

identification of answers to decide which answer will get accepted.

The rest of the paper is organized as follows. We present study on related work in Section II. Section III describes our research questions with research data. The research methodology used is presented in Section IV. The classifier modelling with experimental results has been presented in Section V. Section VI comprises of the extensive feature impact analysis related to our work. The potential threats to validity are discussed in Section VII. Finally, we conclude our work with their future scope in Section VIII.

II. RELATED WORK

''Activeness'', ''Topical interest'', ''Expert Computation'' and ''Recommending the best answer'' are the recent topics on CQA which has attracted many researchers' interest.

In [3], the activeness of users' has been explored in CQA. They have shown how badges and reputation scores are related to find activeness in different forums based on statistical analysis. In [4], the StackOverflow posts has been analyzed through quantitative (statistical data analysis) and qualitative (user interviews) approaches in order to visualize the activity signatures for the success of CQA. In [1], Anusha et al. discussed on clustering the users of StackOverflow into four clusters namely naive, surpassing, experts and out shiners based on characteristics accounting various metrics by using machine learning algorithm in order to predict the users' activities. But they mainly focused on: who dominate the community and how their expertise levels make impact on reputation in the community.

The generative model has been proposed by Guo et al. [5] based on topical interest of the user for recommending answer providers. In [2], Barua et al. has worked on the goal of uncovering topic interest, main discussion topics and technology trends over time with the help of statistical topic modelling [6].

Finding the best answer from a list of answers can be seen as the problem of ranking the answers, where the best answer gets the maximum rank. Ranking technique can also be utilized effectively for selecting the best answer. Some of the popular state-of-the-art ranking techniques are discussed here. The PageRank [7] and HITS algorithm [8] are graph-based approach to rank the web pages. These algorithms consider a web page as a node of the graph and a direct-link from one web page to another as an edge of the graph. The PageRank algorithm computes a random surfer landing probability based on Markov chains of a given web page and accordingly ranks them. The HITS algorithm computes authority and hub value of the web page. Authority value is used to estimate the content of the web page whereas the hub value is used to estimate the links of the web page with other web pages. The ExpertiseRank [9] which is a variant of PageRank [7] computes the expertise score of CQA users. In addition to the graphical features, this algorithm also includes a metric Z-score based on (i) the number of answers given by a user and (ii) the number of questions asked by that user. Their results suggest that a simple metric like Z-score outperforms complex graph based algorithm such as PageRank. In [10], the similar work has been proposed for expert identification in Yahoo! Answers, however they used the number of best answers given by the user as a metric to compute the expertise level of the user based on clustering algorithms. The CQARank [11] algorithm has been proposed to measure user interests and expertise score under different topic using Topic Expertise Model (TEM), which is a novel probabilistic generative model with Gaussian Mixture Model (GMM) hybrid. Zhou et al. [12] proposed the topic sensitive random surfer model (TSPR) by considering the topical similarity among users when setting the affinity weight ignored in TEM [11] for expert finding. Yang and Manandhar [13] learnt the latent feature space of both user and tag to build user-tag matrix and proposed tag based expert recommendation model with the help of probabilistic matrix factorization (PMF) [14], and they showed that the results as well as the computational time obtained by user-tag matrix using PMF outperform those of TEM [11] in the domain of expert recommendation. Pal et al. [15] explored users' question selection preferences through probabilistic model to run machine learning algorithms in order to identify experts and potential experts in CQA.

Treude et al. [16] analyzed StackOverow and categorized the questions using tag and question coding. They used five tag and ten question categories statistically to identify the kind of questions asked and answered. In [17], Wang et al. have extended the work presented in [16] by investigating the distribution of questioners and answerers. They used Latent Dirichlet Allocation (LDA), a well known topic modelling technique [6], to identify topics from questions for learning and assigned a question to several topics with some probabilities. In [18], the User Topical Ability Model (UTAM) has been proposed that exploits both users' expertise and descriptive abilities in CQA sites. UTAM is a probabilistic model, to depict the topic-specific user ability based on textual content (words and tags) and voting scores of questions. In addition to textual and voting information, they also explore social links within a Q&A community by integrating the results of UTAM to describe User Social Topic Ability (USTA). In [19], a topic-sensitive probabilistic model has been proposed that extends the PageRank algorithm [7]. They combined the link information with user information to overcome the drawback of existing link analysis techniques. In [20], Tian et al. have predicted the best answerer for a new question on CQA site considering both topical interest and expertise of the user relevant to the topics of the question asked. They used LDA to identify topical interest from previous answers given by the user, while expertise level is computed using collaborative voting mechanism. In [21], Riahi et al. compared the statistical topic modelling techniques namely LDA [6] and Segmented Topic Model (STM) [22] with the help of traditional information retrieval approaches namely Language Model (LM) [23] and TF-IDF [24]. They found that STM outperforms LDA, LM and TF-IDF.

Tian et al. [25] uses learning from labeled data of question-answer features using classification and predicted the best answer for the question for which the answer is not accepted yet. Similarly, Shah and Pomerantz [26] evaluated and predicted the best answer using classifier learned from the features of question, answer and user in order to meet the satisfaction level given by human under 13 different criteria. In [27], the high quality answers has been identified by running hybrid hierarchy of classifiers trained on the features identified in order to predict the quality score of the answer. They trained type-based quality classifier to aggregate overall quality score of an answer to improve the accuracy which is not achieved by modelling each individual Q&A pair differently.

Motivated from the work presented by Treude et al. [16], this paper addresses the answer to unanswered research question 3 - *how are the best answers selected?*. First, we use activity signatures [1], [3], [10], domain knowledge [13] and topical similarity [2] of the user to identify active answerers in the domain of the questions asked. Second, we use topical interest, topical expertise for expertise computation as used in [11], [20], and [21] using topic modelling and voting scores. Third, we find topic relevancy to find the relationship between Q&A pairs as in [2]. Next, we focus on features involved in posts as in [25] and [26] for predicting whether the answer to the question will be accepted or not based on different classifiers.

III. EMPIRICAL STUDY

The empirical study of the work is performed based on extensive literature survey in terms of research questions. As of now, we didn't find any model from the literature which automatically accepts the best answer. The first question that comes in our mind is that if every question receives multiple answers then how and what basis the best answer is accepted for the question. In order to answer this question, we study StackOverflow CQA site and develop a model to accept the best answer automatically for the asked question which allows us to formulate the following four research questions.

A. RESEARCH QUESTIONS

1) *Who provide answers to the question on stack overflow?* The answerer works on multiple domains and has knowledge in a wide array of topics, tools, technologies and programming languages. StackOverflow CQA site is designed to resolve the challenges faced by the users in computer programming by knowledge sharing in the form of question answering. Clearly, you can find that the question asked by the user must belong to a particular domain of the programming. So, how to find the domain of the question that can help us to identify the answerer in that domain who answers the question [16]. Moreover, the domain knowledge of the answerer is the possible research area that has high impact in expert identification.
2) *How much expertise does the answerer has in (i) same domain (ii) different domain?* Expertise computation [9], [11], [13], [15] in CQA has been gaining popularity amongst researcher over recent years. Answerer of the question may have expertise on the same domain as well as on different domain of the question. In general, the answer given by the answerer of the same domain has high quality and more likely to be accepted. In addition to the domain knowledge, the number of question asked and the number of answer given in various domains play important roles for computing the expertise level.
3) *To what extent the answer is relevant to the question?* The asker or community, review the received answers so as to meet the requirement of the asker described by the question. We found that the requirement of the question is explained through textual information. We use the textual information of both question and answer to find the closeness between them. The more similarity between textual information of question and answer represents the more relationship [2]. Text mining is the area to deal the textual content of the document to learn the properties for further inference.
4) *How the best answer is selected amongst a set of answers?* There is no limitation in posting the number of answers to the question. The asker or community then review the answers manually to accept the answer that explains the requirement of the question. It takes a lot of time to review these answers as the size of StackOverflow growing rapidly. This would encourage us to work on this field.

B. DATASET DESCRIPTION

StackOverflow is an interactive CQA site for exchanging the knowledge in the software engineering field. It provides the wide variety of functionality for users to gain the knowledge. The basic function is asking and answering the questions by the registered user. The users are free to give their opinion about questions/answers either in terms of up-voting/down-voting or posting comments on them. The users gain the reputation through different activities like answering the question, their answer is accepted etc. The access privileges on the site are decided by the reputation score of the users, which indicates how useful the user is for the community. Similarly, the importance of the post (question/answer) is determined using the score obtained by the number of up-votes minus the number of down-votes for the post. Furthermore, the asker or community review the answers manually to select an answer as the accepted answer, which satisfies the requirement of the asker and represents that it is the best answer for the given question. Table 1 shows the general statistics of the StackOverflow dataset till May 31, 2015.

StackOverflow offers its data publicly available through Stack Exchange Data Explorer and XML format data dump under creative common licence. We have used the complete dataset about question posts created between

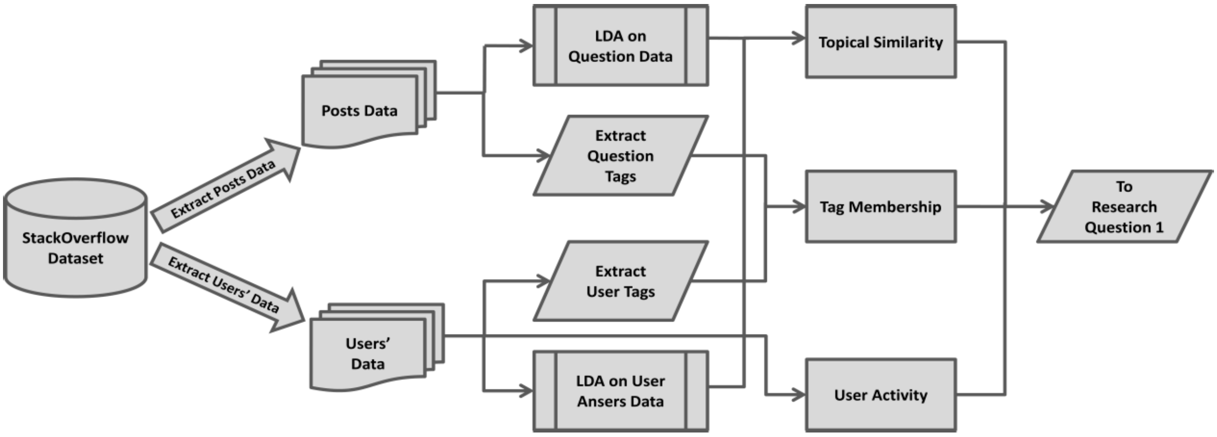


FIGURE 1. Research Methodology for RQ 1.

TABLE 1. General statistics of the stackoverflow.

Description	Count
Question with an Accepted Answer	54,00,897
Question without an Accepted Answer	40,04,446
Question not answered	10,10,175
Total Question	1,04,15,518
Total Answer	1,58,03134
Asker	17,46,595
Answerer	11,42,761
Commenter	14,22,041
Total Registered User	42,86,030
Tag	40,871

TABLE 2. Extracted dataset statistics.

Description	Count
Question	3,185
Answer	14,384
Asker	3002
Answerer	10340
Tags	1480

January 01, 2015 and March 31, 2015 on Stack Overflow through Stack Exchange Data Explorer. Additionally, we used 3,23,972 prior answers given by 10,340 answerer involved in the extracted dataset to judge topical expertise. The statistics about the dataset relevant to our study has been presented in Table 2. However, we have not included the data of posts (i) which have less than five answers (ii) for which answer is not accepted yet so as to analyze how the answers will get accepted.

IV. RESEARCH METHODOLOGY

A. RQ 1. (WHO ANSWERS THE QUESTION ON STACK OVERFLOW?)

In StackOverflow, there are many questions from various topics related to programming. There are about 10M questions, 17M answers, 4.5M users and 42K tags in the StackOverflow till May 31, 2015. So, it's a challenge to detect who answers the questions of the StackOverflow. In our analysis, we found that the user who belongs to the domain of the question and have prior knowledge in that domain can answer the question on StackOverflow. Our criteria for evaluating the user's knowledge of the domain are based on user's tags and topics on which the user has been previously involved. But knowledge alone is not sufficient for a user to answer on a particular question and we found that there is a strong relation of activeness of the particular user during the period when the question was asked. The detailed research methodology for RQ 1 is presented in Figure 1. Based on our study, we try to identify answerer based on three key features:

1) TAG MEMBERSHIP

It is defined as the association amongst tags of the question with the tags of the user on which s/he has answered previously across StackOverflow and given by the following formula.

$$TagMembership\ (\mu_{TM}) = \frac{\emptyset\,(t,u) \cap \emptyset\,(t,q)}{\emptyset\,(t,q)} \tag{1}$$

where, $\emptyset\,(t,u)$ and $\emptyset(t,q)$ gives the tags associated with user (u) and question (q) respectively.

2) TOPICAL SIMILARITY

It is defined as a cosine similarity between the topic terms of the question and previous answers given by the answerer and given by the following formula.

$$TopicalSimilarity\ \left(TSim_{qa_i}\right) = \frac{|Tt\ (q) \cap Tt\ (a_i)|}{\sqrt{|Tt\ (q)| * |Tt\ (a_i)|}} \tag{2}$$

where, $Tt\ (q)$ and $Tt\ (a_i)$ are the set of topic terms of question (q) and prior answers (a_i) given by the answerer respectively.

3) ACTIVENESS

The average number of answers given by the answerer per day and it is given by the following formula.

$$Activeness\ (A) = \frac{|a_i|}{N} \tag{3}$$

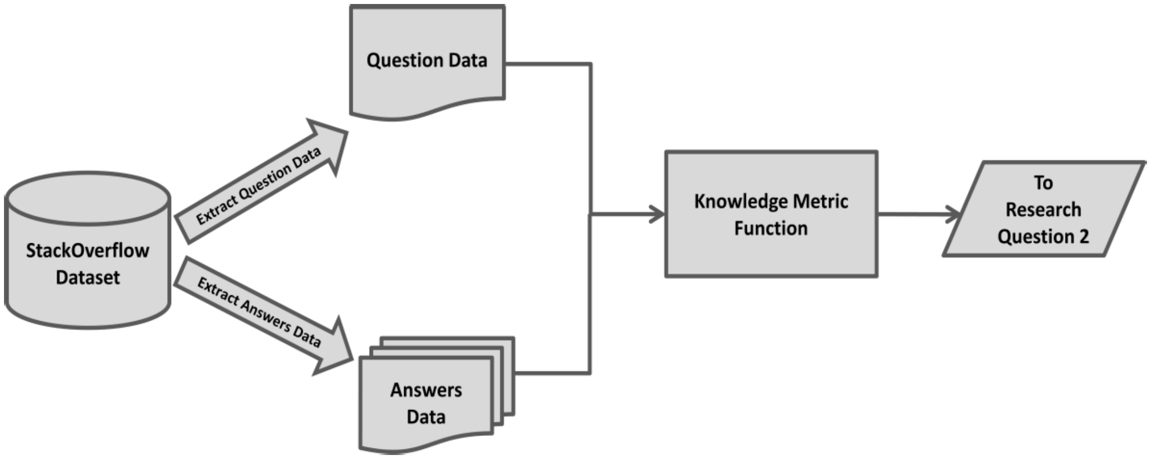


FIGURE 2. Research Methodology for RQ 2.

TABLE 3. Sample data (PostID = 27729802) for RQ 1.

Answer (AnswerID)	Answerer (AnswerID)	Tag Membership (μ_{TM})	Topical Similarity ($TSim_{qa_i}$)	Activeness (A)
27730172	4299161	0.5	0.73	0.29
27730248	483408	0.5	0.81	0.09
27730892	**73226**	**0.5**	**0.87**	**1.82**
27731076	2964963	0.5	0.82	0.61
27731845	137508	0.5	0.83	0.35

where, $|a_i|$ is the number of answers given by the answerer in N days.

From Table 3, we can draw the inference that the active user, who has knowledge on domain of the question, can answer the question. We use Pearson's Correlations between the number of answer given by the user and features of RQ 1 to see the effect in identifying the answerer of the question. Pearson's correlation coefficient (r) ranges from -1 to $+1$ and p represents the significant level. The positive value of (r) indicates both variables are increases or decreases together i.e. positive correlation, whereas negative value of (r) indicates that as one variable increases, so the other decreases, and vice versa i.e. negative correlation. For the strength of correlation, $|r| < 0.3$ indicates small correlation, $0.3 \leq |r| < 0.5$ indicates medium correlation, and $|r| > 0.5$ indicates strong correlation. First, the Tag Membership ($r = 0.472, p < 0.001$) and the Topical Similarity ($r = 0.458, p < 0.001$) are positively correlated with the number of answer given by the users, indicating that the higher value of Tag Membership and Topical Similarity increases the number of answer given by the users. Second, Activeness is by default positively correlated as per the definition.

B. RQ 2. (HOW MUCH EXPERTISE DOES THE ANSWERER HAS IN (i) SAME DOMAIN (ii) DIFFERENT DOMAIN OF THE QUESTION?)

We observed that the number of answerer gives the answer to the question; these answerers have certain level of knowledge in the question domain. The knowledge of the answerer can be derived from the metrics such as number of up-votes, reputation and percentage of accepted answer as in Figure 2. As the knowledge of the answerer increases the expertise level also increases, consequently the respective answers are likely to be accepted.

Now, we focus on the expertise of the answerer by using the knowledge level metrics in two categories as follows.

1) TAG SCORE

The number of up-votes the answerer has on the tags to which s/he has given the prior answer and is given by the following formula.

a: SAME DOMAIN

$$TagScore\ \ (TS_{sd}) = \sum_{i=1}^{|\emptyset(t,q)|} UV_i \tag{4}$$

b: DIFFERENT DOMAIN

$$TagScore\ \ (TS_{dd}) = \sum_{i=1}^{|\emptyset(t,A)|} UV_i - \sum_{i=1}^{|\emptyset(t,q)|} UV_i \tag{5}$$

where, $|\emptyset\,(t, q)\,|$ is the number of common tags between question and answerer, $|\emptyset\,(t, A)\,|$ is number of tags associated with the answerer, and UV_i is the number of up-votes of the answerer in i^{th} tag.

2) TOPIC REPUTATION

The prior contribution of the topics of question on the overall reputation (R) of the answerer and is given by the following formula.

a: SAME TOPIC

$$TopicReputation\ \ (TR_{st}) = TSim_{qa_i} * R \tag{6}$$

TABLE 4. Sample data (PostID = 27729802) for RQ 2.

Answerer (AnswererID)	Tag Score (SD)	Tag Score (DD)	Topic Reputation (ST)	Topic Reputation (DT)	Accepted Answer (SD)	Accepted Answer (DD)	Overall Reputation
4299161	2	3	111.6	40.4	0	0	152
483408	98	179	3166.1	447.9	30	332	3614
73226	**11545**	**16849**	**168529.8**	**39162.2**	**1334**	**11846**	**207692**
2964963	52	110	2435.2	543.8	44	432	2979
137508	124	518	6177.1	1237.9	20	618	7415

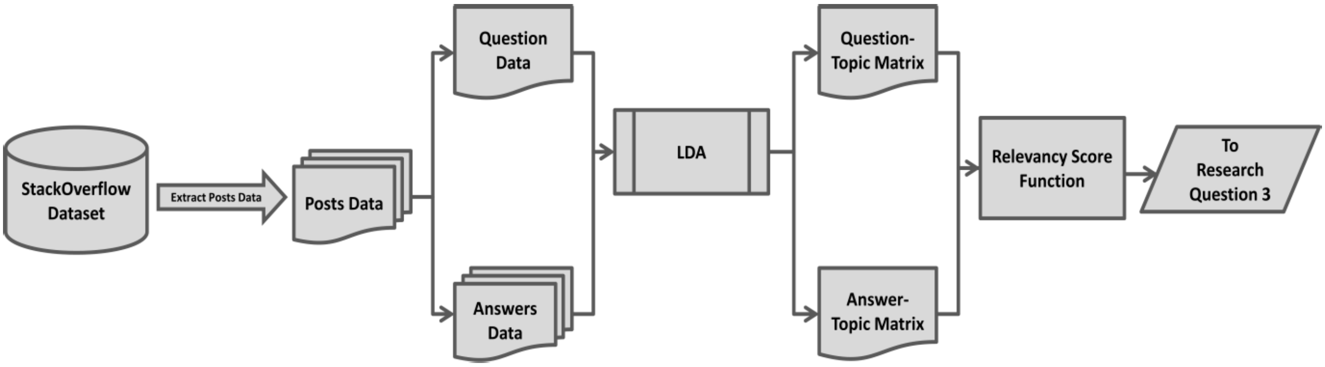


FIGURE 3. Research Methodology for RQ 3.

b: DIFFERENT TOPIC

$$TopicReputation\ \ (TR_{dt}) = \left(1 - TSim_{qa_i}\right) * R \qquad (7)$$

where, R is the reputation earned by the answerer and $TSim_{qa_i}$ is the topical similarity between question and prior answers given by the answerer.

3) ACCEPTED ANSWER

The ratio of number of accepted answer to the total number of answer given by the answerer and is given by the following formula.

a: SAME DOMAIN

$$AcceptedAnswer\ \ (AA_{sd}) = \frac{\{|aa_i|\ \ Such\ That\ \mu_{TM} \geq 0.2,\ for\ all\ i\}}{\{|a_t|\ \ Such\ That\ \mu_{TM} \geq 0.2,\ for\ all\ t\}} \qquad (8)$$

b: DIFFERENT DOMAIN

$$AcceptedAnswer\ \ (AA_{dd}) = \frac{\{|aa_i|\ \ Such\ That\ \mu_{TM} = 0,\ for\ all\ i\}}{\{|a_t|\ \ Such\ That\ \mu_{TM} = 0,\ for\ all\ t\}} \qquad (9)$$

where, $|aa_i|$ is the number of accepted answer, $|a_t|$ is the total number of answer, and μ_{TM} is the tag membership of answerer to the question.

The above three metrics along with the overall reputation is combined to explore the expertise of the answerer posses on the (i) Same Domain (ii) Different domain. From Table 4, we can draw the inference that not all answerer are experts on the question domain but certainly they have knowledge. Clearly, answerer 73226 is expert in the question domain; while other answerer has less expertise or knowledge in question domain but they may be experts in different domain except the answerer 4299161.

Now, we use Pearson's Correlations between the Z-score [9] and metrics of RQ 2 to see their effect in describing the expertise level. The Tag Score (SD) ($r = 0.637, p < 0.001$), Tag Score (DD) ($r = 0.598, p < 0.001$), Topical Reputation (ST) ($r = 0.87, p < 0.001$), Topical Reputation (DT) ($r = 0.757, p < 0.001$), Overall Reputation (R) ($r = 0.875, p < 0.001$), Accepted Answer Ratio (SD) ($r = 0.774, p < 0.001$), and Accepted Answer Ratio (DD) ($r = 0.74, p < 0.001$) are positively correlated with the Z-score of the answerers, indicating that the higher value of metrics represents higher expertise in the respective domain.

C. RQ 3. (TO WHAT EXTENT THE ANSWER IS RELEVANT TO THE QUESTION ASKED?)

In RQ 2, we found the knowledge level of each answerer in terms of their expertise they have, but it is not always true that the answer given by the answerer is that much relevant to the question as required. As the question and answer is composed of various topics, we try to find the compatibility of the answer with the question using various topics to meet the satisfaction level of the asker as in Figure 3 [2].

Now, we calculate the relevancy of the answer given by each answerer to the question asked and is given by the following formula.

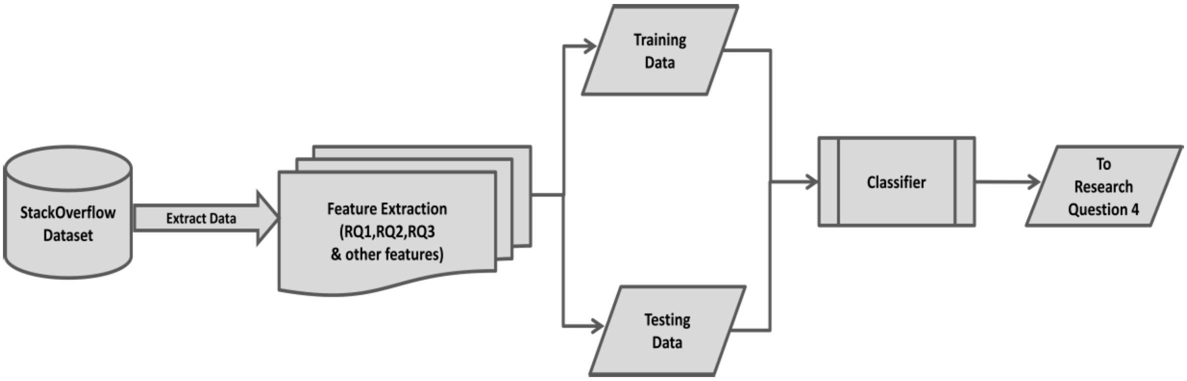


FIGURE 4. Research Methodology for RQ 4.

1) TOPIC RELEVANCY

It is defined as the relationship between topics found in questions (Zq_i) and topics found in the corresponding answers (Za_j) and is given by the following formula.

$$TopicRelevancy\ (TR_{qa}) = \sum_{\substack{d_q \in Q, d_a = a(d_q) \\ \mu(d_q, Zq_i) \geq \delta \\ \mu(d_a, Za_j) \geq \delta \\ \forall i,j}} \mu\left(d_q, Zq_i\right) * \mu\left(d_a, Za_j\right) \quad (10)$$

TABLE 5. Sample data (PostID = 27729802) for RQ 3.

Answers (AnswerID)	Topic Relevancy (TR_{qa})
27730172	0.066
27730248	0.111
27730892	**0.135**
27731076	0.107
27731845	0.124

where, Q is the set of questions posted, d_q is the textual information of the question q, A is the set of answers posted, d_a is the textual information of the answer a and is equals to $A(d_q)$, Zq_i is the i^{th} topic of the question q, Za_j is the j^{th} topic of the answer a, $\mu\left(d_q, Zq_i\right)$ is membership value of i^{th} topic of the question Zq_i, and $\mu\left(d_a, Za_j\right)$ is membership value of j^{th} topic of the answer Za_j. To limit the number of topics, we set threshold value $\delta = 0.1$ for both question and answer. Table 5 represents the relationship between question and answer. The higher the value of TR_{qa} represents the higher relevancy between question and their corresponding answer. The relevancy between question (PostID = 27729802) and answer (PostID/AnswerID = 27730892) is higher than other QA pairs. Pearson's Correlations is used between the *answer score* and metric of RQ 3 to see the effect and found ($r = 0.443, p < 0.001$), which represents *topic relevancy* is positively correlated with the *answer score* of the answer, indicating that the higher value of metric represents high relevancy between question and answer.

D. RQ 4. (HOW THE BEST ANSWER IS ACCEPTED AMONGST A SET OF ANSWERS?)

Currently, the evaluation of the answers is carried out manually for accepting the answer that meet the requirement of the asker, which is explained through the question. Generally, the answer is evaluated using evaluation parameters, these evaluation parameters are hidden in the current scenario. We compile the answer of above three research questions, which acts as the baseline in order to answer the research question RQ 4 along with one more metric i.e. *time span* of the answer using classifier. The detailed research methodology is shown in Figure 4.

1) TIME SPAN OF ANSWER

It is defined as the time difference between question posting and corresponding answer posting and can be expressed as:

$$TimeSpan\ \ (TS_a) = t\ (a \rightarrow q) - t\ (q) \quad (11)$$

where, $t(a \rightarrow q)$ is the answer a posting time to the question q, and $t(q)$ is question posting time.

TABLE 6. Sample data (PostID = 27729802) for RQ 4.

Answers (AnswerID)	Time Span (min) (TS_a)
27730172	57
27730248	71
27730892	**159**
27731076	185
27731845	297

Table 6 gives elapsed time of the answers after posting the corresponding question of PostID 27729802. The asker is expecting the answer as early as possible so as to meet the requirement. Therefore, the time span of the answer is very important metric in accepting the answer.

Again, we use Pearson's Correlations between the *time span* and the class whether the answer is accepted or not to see the effect and found ($r = -0.233, p < 0.001$), which represents *time span* is negatively correlated with

TABLE 7. Prediction results using classifiers.

Classifier	Class	Confusion Matrix		ROC-AUC	PRC-AUC	Gmean	Accuracy (%)
BayesNet	Accepted	1452	1733	**0.654**	**0.286**	**0.582**	45.59
	Non-Accepted	3703	10681		**0.885**		74.26
	Weighted Average	------	------		**0.776**		69.06
NaiveBayes	Accepted	1302	1883	0.642	0.283	0.568	40.88
	Non-Accepted	3033	11351		0.874		78.91
	Weighted Average	------	------		0.766		72.02

TABLE 8. Characteristics of actual dataset.

S. No.	Features	Accepted Answer			Non-Accepted Answer		
		Median	Mean	SD	Median	Mean	SD
1.	Tag Membership (μ_{TM})	1	0.77	0.32	0.75	0.67	0.36
2.	Topical Similarity $(TSim_{qa_i})$	0.82	0.55	0.41	0.75	0.44	0.43
3.	Activeness (A)	0.26	1.01	1.92	0.12	0.66	1.53
4.	TSSD (TS_{sd})	73	1865	8693	16	647	3448
5.	TSDD (TS_{dd})	140	3484	18767	33	1157	7712
6.	TRST (TR_{st})	2194	27934	73276	80	11741	40691
7.	TRDT (TR_{dt})	1138	5591	14579	441	2633	7657
8.	AASD (AA_{sd})	23	290	896	5	116	454
9.	AADD (AA_{dd})	299	3460	12312	77	1373	5922
10.	Reputation (R)	3829	33526	85012	1203	14375	47289
11.	Topic Relevancy (TR_{qa})	0.099	0.1	0.02	0.098	0.09	0.02
12.	Time Span (TS_a)	12	2358	13567	13	3491	18974

TABLE 9. Characteristics of normalized dataset.

S. No.	Features	Accepted Answer			Non-Accepted Answer		
		Median	Mean	SD	Median	Mean	SD
1.	Tag Membership (μ_{TM})	1	0.59	0.46	0.5	0.48	0.46
2.	Topical Similarity $(TSim_{qa_i})$	0.91	0.58	0.46	0.003	0.46	0.47
3.	Activeness (A)	0.22	0.41	0.41	0.08	0.29	0.38
4.	TSSD (TS_{sd})	0.21	0.43	0.44	0.02	0.24	0.37
5.	TSDD (TS_{dd})	0.23	0.43	0.43	0.03	0.24	0.37
6.	TRST (TR_{st})	0.12	0.39	0.44	0	0.22	0.37
7.	TRDT (TR_{dt})	0.31	0.46	0.42	0.07	0.28	0.37
8.	AASD (AA_{sd})	0.22	0.44	0.44	0.02	0.24	0.37
9.	AADD (AA_{dd})	0.25	0.45	0.43	0.04	0.25	0.37
10.	Reputation (R)	0.28	0.46	0.43	0.05	0.26	0.37
11.	Topic Relevancy (TR_{qa})	0.46	0.48	0.37	0.44	0.42	0.36
12.	Time Span (TS_a)	0.13	0.33	0.39	0.17	0.36	0.4

the *class* of the answer, indicating that the higher value of metric represents less chance of acceptance of the answer.

V. CLASSIFIER MODELING AND EXPERIMENTAL RESULTS

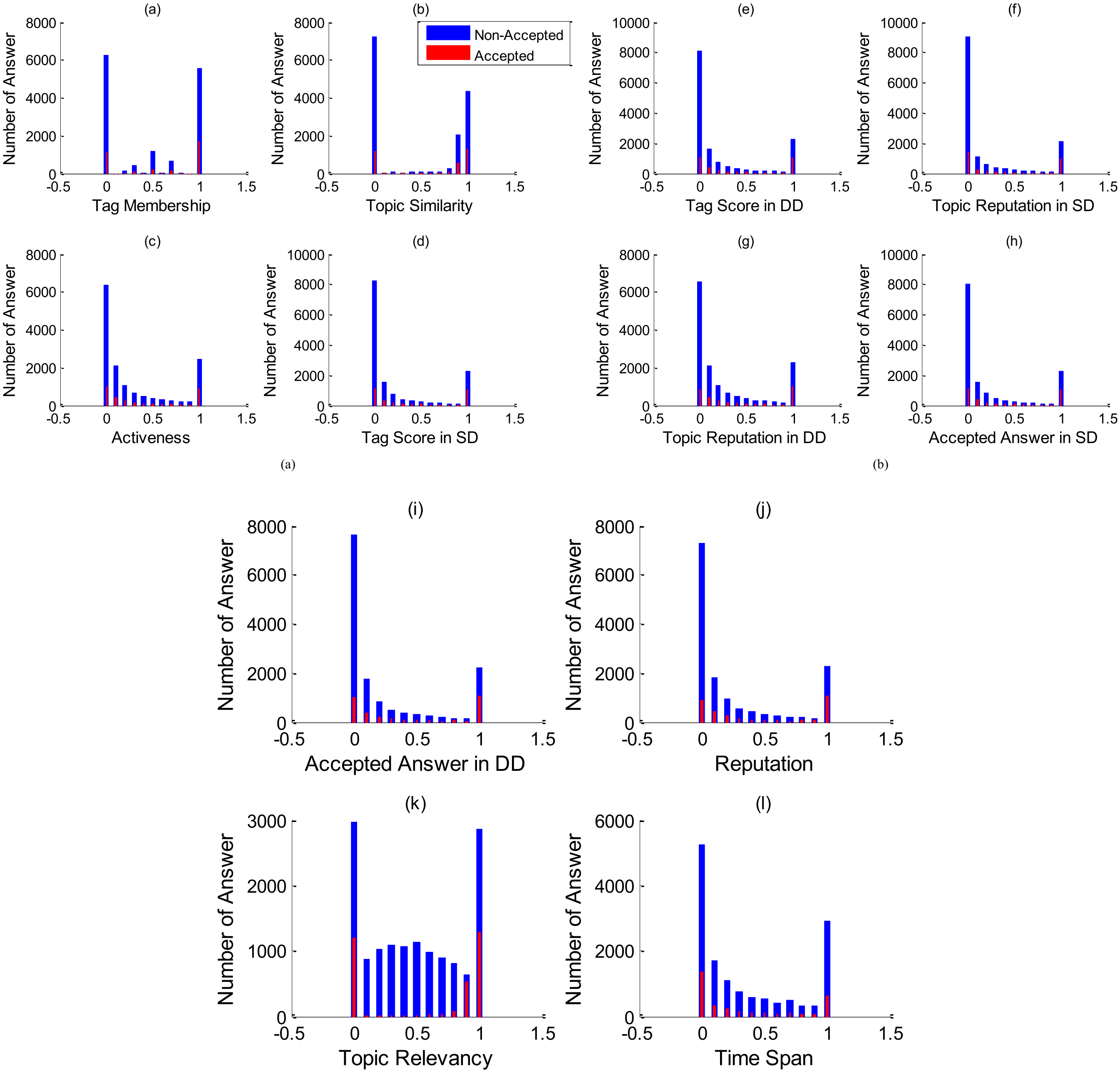
A. CLASSIFIER MODELING

Based on a posts feature vector that we have extracted, the answer is classified whether accepted or not. We used *BayesNet and NaiveBayes* for the task of binary classification, in which the generative model is utilized for modelling question and answer based on Gaussian distribution. The model assumes that answers are Gaussian distributed in terms of their acceptance. The scatter plot of features for the dataset can be drawn and the contours of the Gaussian distribution fitted based on the MLE estimates. The smaller contours will capture majority of the accepted answer and some of the non-accepted answers as well. The model parameters of Gaussian distribution are $\theta = \{\mu, \Sigma$.

$$P(x|\theta) = \frac{1}{2\pi^{\frac{|x|-1}{2}}\,|\Sigma|^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}} \quad (12)$$

The model parameters for the two classes are $\theta_{YES}\mu_{YES}$, Σ_{YES} and $\theta_{NO}\mu_{NO}$, Σ_{NO}. We assume that an answer acceptance is independent and identically distributed (i.i.d.) which simplifies Maximum Likelihood Estimation (MLE):

$$\theta^{MLE} = arg\ max_\theta \{P(D|\theta)\} = arg\ max_\theta \left\{\prod_{a_q} P\left(x_{a_q}|\theta\right)\right\} \quad (13)$$



(a)

(b)

(c)

FIGURE 5. **(a). Statistical distribution (a-d) of the features. (b): Statistical distribution (e-h) of the features. (c). Statistical distribution (i-l) of the features.**

For classification, Bayes rule is used to generate the posterior distribution of class conditioned on feature vector:

$$P\left(YES|a_q\right) = P\left(x_{a_q}|\theta_{YES}^{MLE}\right) \cdot P\left(YES\right) \qquad (14)$$

where, $P(YES)$ is prior probability of an answer being accepted. Prior probability is the ratio of accepted answers to total answer in the training data. We also compute posterior probability of an answer belonging to the other class, whichever class has a higher probability.

B. EXPERIMENTAL RESULTS

We model the features of a post in Bayes rule framework for evaluating the performance of binary classification. *BayesNet* and *NaiveBayes* classifiers are used in the prediction task and results are presented in Table 7. The well known 10 fold cross validation is used for training and testing of the dataset. We run the Weka implementation of the classifiers with default settings [28].

To assess the prediction quality of the classifier, we use four evaluation measures namely Receiver Operating

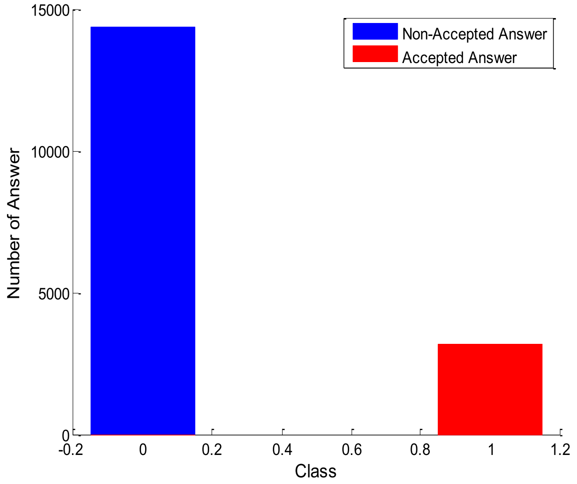


FIGURE 6. Statistical distribution of the Class.

Characteristics Area Under Curve (ROC-AUC), Precision Recall Area Under Curve (PRC-AUC), G-mean and accuracy. First three evaluation measures are generally used to assess the performance of the classification where the dataset is imbalanced as we have. Table 7 presents the experimental results and indicating that *BayesNet* classifier outperforms over *NaiveBayes* classifier.

VI. FEATURE IMPACT ANALYSIS

A. STATISTICAL ANALYSIS

The characteristics of Accepted and Non-Accepted are investigated through the various measures like central tendencies (median and mean) and standard deviations of the extracted features. Table 8 and Table 9 shows that the characteristics of the actual and normalized dataset respectively. The range value (min and max) of the features doesn't give any idea about acceptance of the answer. It happens because of the large proportions of Non-Accepted answer compared to Accepted answer. However, It is observed that the median and mean value of the features for Accepted answer are clearly greater than that of Non-Accepted answer, which reveals that the answers having greater value for the features from S. No. 1 to 11 and less value for the feature of S. No. 12 are more likely to be accepted.

B. STATISTICAL DISTRIBUTION

In order to see the distribution of the features amongst both the classes Accepted and Non-Accepted, the overlaying bar graph has been presented for these features of normalized dataset in Figure 5(a), Figure 5(b) and Figure 5(c). On average, the number of Accepted answer are increasing from (0, 1] and the number of Non-Accepted answer are decreasing from [0, 1) in x-axis for all the features except the last feature i.e. time span, which is opposite in nature w.r.t. to other features. This indicates that the features have great importance based on their values (normalized) in the class distribution shown in Figure 6.

C. FEATURE RANKING

The statistical measures and statistical distribution of the features are different for both Accepted and Non-Accepted class presented in previous section. This finding indicates that the features are key in predicting whether a question will get accepted or not. However, the previous section doesn't deal with the relative importance of the features for differentiating Accepted and Non-Accepted answers. Therefore, we use three well known ranking algorithms for feature evaluation based on information gain, gain ratio and chi-squared statistics to rank the features. A feature ranking algorithm is helpful to filter unimportant features in the prediction task.

1) *Information Gain:* In information theory, the information gain of a random variable is defined as the change in information entropy from a prior state to a state that takes the variable as given. Therefore, the information gain of a particular feature in classifying if a question is Accepted or Non-Accepted is:

$$InfoGain\,(C, f_i) = H\,(C) - H\left(\frac{C}{f_i}\right) \tag{15}$$

where, C represents a particular class (Accepted or Non-Accepted), f_i denotes the feature, and H denotes information entropy.

2) *Gain Ratio:* Although information gain is usually a good measure for deciding the relevance of a feature, it favours the features that can take on a large number of distinct values. Therefore, we have used gain ratio to rank our features, which overcomes the previous problem. Gain ratio is defined as:

$$GainRatio\,(C, f_i) = \frac{\left(H\,(C) - H\left(\frac{C}{f_i}\right)\right)}{H\,(f_i)} \tag{16}$$

3) χ^2 *Statistic:* Feature Selection using Chi-square $\left(\chi^2\right)$ test is yet another and very commonly used method. Attribute evaluation via chi-squared evaluates the importance of a feature by computing chi-squared statistic value with respect to the class. The null hypothesis H_0 assumes that the feature and class are independent based on their occurrences. We then rank the features by chi-squared formula:

$$\chi^2\,(D, f_i, C) = \sum_{e_{f_i}\in\{0,1\}} \sum_{e_c\in\{0,1\}} \frac{\left(O_{e_{f_i}e_c} - E_{e_{f_i}e_c}\right)^2}{E_{e_{f_i}e_c}} \tag{17}$$

where, $O_{e_{f_i}e_c}$ is the observed frequency and $E_{e_{f_i}e_c}$ is the expected frequency in dataset D for the feature f_i and class C. e_{f_i} and e_c corresponds to the value of the feature f_i and class C, if the queried value is present then e_{f_i} and e_c will be 1 otherwise 0. The greater the value of χ^2 statistic the greater the importance of the feature is against H_0.

However, gain ratio gives an extra benefit to the features with very low information gain. Therefore, we use above three

TABLE 10. Feature ranking.

Feature Rank	Features	Information Gain	Features	Gain Ratio	Features	Chi-Squared Statistic
1.	Reputation (R)	0.03155	R	0.01592	R	772.5251
2.	AADD (AA_{dd})	0.02966	AA_{dd}	0.01492	AA_{dd}	737.1486
3.	TSSD (TS_{sd})	0.02901	TS_{sd}	0.01482	TS_{sd}	722.4403
4.	AASD (AA_{sd})	0.02813	TR_{st}	0.01478	AA_{sd}	715.9528
5.	TSDD (TS_{dd})	0.02788	AA_{sd}	0.01433	TS_{dd}	695.6157
6.	TRDT (TR_{dt})	0.02684	TS_{dd}	0.01427	TR_{dt}	653.464
7.	TRST (TR_{st})	0.02029	TR_{dt}	0.01359	TR_{st}	527.1309
8.	Activeness (A)	0.0112	μ_{TM}	0.00847	A	278.3358
9.	Tag Membership (μ_{TM})	0.00829	A	0.00751	μ_{TM}	204.6413
10.	Topical Similarity $\left(TSim_{qa_i}\right)$	0.00717	$TSim_{qa_i}$	0.00718	$TSim_{qa_i}$	172.5858
11.	Topic Relevancy $\left(TR_{qa}\right)$	0.00512	TR_{qa}	0.00629	TR_{qa}	169.2828
12.	Time Span (TS_a)	0.00191	TS_a	0.00239	TS_a	48.0421

rankings to select the prominent features. We use the Weka implementation [28] of Information Gain, Gain Ratio and Chi-squared Ranking algorithm with default settings to rank the features defined as above. The detailed ranking result with their corresponding values is presented in Table 10. The rankings are same for information gain and chi-squared whereas there are some differences in ranking using Gain Ratio. But in total, the three ranking algorithms agreed on selecting the 12 features listed in the table for the proposed problem. From all the rankings, we found that the first 7 features which represent the expertise of the answerer are the most dominant features in predicting whether an answer for the particular question will get Accepted or Non-Accepted. The rest features are also important as discussed in Section IV.

VII. THREATS TO VALIDITY

Although the extensive empirical analysis and experiments has been performed in the evaluation and selection of the features for accepting the answer for the asked question. There are some threats which can affect the results: (i) posts selected for experiments can affect the parameters of studies (ii) Other CQA sites can arrange the metadata in different format, so preprocessing can affect the performance parameters (iii) our model fails to calculate the expertise level for potential experts [15].

VIII. CONCLUSION AND FUTURE SCOPE

In this paper, we study and analyse the StackOverflow answers with their question to predict whether the answer will get accepted or not. We perform an extensive empirical analysis on StackOverflow dataset to answer the four research questions. To answer the research questions, we extracted the corresponding features to answer the framed research questions. Our findings will suggest that: i) prior involvement of the answerer on question tags and topics increases the chance to give the answer for that question ii) expertise will increases the chance in acceptance of the answer iii) topical compatibility between the question and answer increases the satisfaction of asker or community with that answer. We use various statistical methods in order to answer the first three research questions whereas classifier model is used in order to answer the fourth research question based on the previous research questions. With this, we conduct the experiments and found that the extracted features have the great importance for the proposed problem using various feature evaluation metrics. Armed with this observation, we next use classification algorithms based on Bayes rule to predict acceptability of the answer by the asker or community. The evaluation parameter of the classifier reveals that the results are remarkable in predicting the answer acceptability.

As of now, we are the first to analyze and conduct the experiments to predict the acceptance of the answer whatsoever. We have achieved the maximum accuracy of 46% for minority class (accepted), 74% for majority class (non-accepted) and overall accuracy 69% with 0.645 PRC area using *BayesNet* Classifier. The accuracy may be improved in future for both the classes with increased PRC area in several ways. First, classifier model must be designed in such a way that the class imbalance problem doesn't influence the accuracy of both the classes and so overall accuracy keeping in view PRC area. Second, Feature extraction plays a vital role in classification problem, so identifying additional features especially temporal features w.r.t. questions, answers and users can improve the performance of the classifier. Third, Ensemble learning can be devised to model the suggested problem in order to improve the accuracy.

REFERENCES

[1] J. Anusha, V. S. Rekha, and P. B. Sivakumar, "A machine learning approach to cluster the users of stack overflow forum," in *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*. India: Springer, 2015, pp. 411–418.

[2] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? An analysis of topics and trends in stack overflow," *Empirical Softw. Eng.*, vol. 19, no. 3, pp. 619–654, 2014.

[3] V. S. Sinha, S. Mani, and M. Gupta, "Exploring activeness of users in QA forums," in *Proc. 10th Work. Conf. Mining Softw. Repositories*, 2013, pp. 77–80.

[4] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&a site in the west," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 2857–2866.

[5] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 921–930.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," 1999.

[8] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[9] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 221–230.

[10] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: The case of Yahoo! answers," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 866–874.

[11] L. Yang *et al.*, "CQArank: Jointly model topics and expertise in community question answering," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 99–108.

[12] G. Zhou, S. Lai, K. Liu, and J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1662–1666.

[13] B. Yang and S. Manandhar, "Tag-based expert recommendation in community question answering," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 960–963.

[14] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. NIPS*, 2011, pp. 1–8.

[15] 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, 2012, Art. no. 10.

[16] C. Treude, O. Barzilay, and M.-A. Storey, "How do programmers ask and answer questions on the Web? (NIER track)," in *Proc. 33rd Int. Conf. Softw. Eng. (ICSE)*, 2011, pp. 804–807.

[17] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in StackOverflow," in *Proc. 28th Annu. ACM Symp. Appl. Comput.*, 2013, pp. 1019–1024.

[18] B. Yang and S. Manandhar, "Exploring user expertise and descriptive ability in community question answering," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 320–327.

[19] G. Zhou, J. Zhao, T. He, and W. Wu, "An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities," *Knowl.-Based Syst.*, vol. 66, pp. 136–145, Aug. 2014.

[20] Y. Tian, P. S. Kochhar, E.-P. Lim, F. Zhu, and D. Lo, "Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting," in *Proc. Workshops Int. Conf. Social Informat.*, 2013, pp. 55–68.

[21] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 791–798.

[22] L. Du, W. Buntine, and H. Jin, "A segmented topic model based on the two-parameter Poisson–Dirichlet process," *Mach. Learn.*, vol. 81, no. 1, pp. 5–19, 2010.

[23] D. Petkova and W. B. Croft, "Hierarchical language models for expert finding in enterprise corpora," *Int. J. Artif. Intell. Tools*, vol. 17, no. 1, pp. 5–18, 2008.

[24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

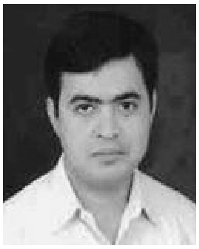
[25] Q. Tian, P. Zhang, and B. Li, "Towards predicting the best answers in community-based question-answering services," in *Proc. ICWSM*, 2013, pp. 725–728.

[26] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 411–418.

[27] H. Toba, Z.-Y. Ming, M. Adriani, and T.-S. Chua, "Discovering high quality answers in community question answering archives using a hierarchy of classifiers," *Inf. Sci.*, vol. 261, pp. 101–115, Mar. 2014.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

TIRATH PRASAD SAHU was born in Baikunthpur, India in 1987. He received the B.Tech. degree in information technology from NIT Raipur, India, in 2009, and the M.Tech. degree in computer science and engineering from SATI, Vidisha, India, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering. He is currently an Assistant Professor with the Department of Information Technology, NIT Raipur. His research interests include data mining, text analytics, and image processing.



NARESH KUMAR NAGWANI received the degree in computer science and engineering from G. G. Central University, Bilaspur, in 2001, the M.Tech. degree in information technology from the ABV-Indian Institute of Information Technology, Gwalior, in 2005, and the Ph.D. degree in computer science and engineering from the National Institute of Technology Raipur, India, in 2013. He was a Software Developer and the Team Lead with Persistent Systems Limited. He is currently an Assistant Professor and the Head of the Department, Computer Science and Engineering, NIT Raipur. He has authored over 40 research papers in various journals and conferences in the field of data mining, text analytics, and software engineering.

SHRISH VERMA received the degree in electronics and telecommunication engineering and the M.Tech. degree in computer engineering from IIT Kharagpur, and the Ph.D. degree in engineering from Pt. Ravi Shankar Shukla University Raipur. He is currently the Head and a Professor with the Department of Electronics and Telecommunication Engineering, National Institute of Technology, Raipur. He has authored over 50 research papers in various journals and conferences in the field of computer and communication networks, distributed processing, data mining and analysis, text analytics and software engineering.