Egyptian Informatics Journal 19 (2018) 21–31

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

A hybrid model to predict best answers in question answering communities

Dalia Elalfy [*], Walaa Gad, Rasha Ismail

Information Systems Department, Faculty of Computer & Information Sciences, Ain Shams University, Abbasia, Cairo, Egypt

ARTICLE INFO

Article history:
Received 22 September 2016
Revised 13 March 2017
Accepted 22 June 2017
Available online 29 June 2017

Keywords:
Question answering communities
Knowledge exchange
Expert
Best answer
Feature extraction
Content feature
Non-content feature
Hybrid classifier

ABSTRACT

Question answering communities (QAC) are nowadays becoming widely used due to the huge facilities and flow of information that it provides. These communities target is to share and exchange knowledge between users. Through asking and answering questions under large number of categories.

Unfortunately there are a lot of issues existing that made knowledge process difficult. One of those issues is that not every asker has the knowledge and ability to select the best answer for his question, or even selecting the best answer based on subjective matters. Our analysis in this paper is conducted on stack overflow community. We proposed a hybrid model for predicting the best answer. The proposed model is consisting of two modules. The first module is the content feature which consists of three types of features; question-answer features, answer content features, and answer-answer features. In the second module we examine the use of non-content feature in predicting best answers by using novel reputation score function. Then we merge both of content and non-content features and use them in prediction. We conducted experiments to train three different classifiers using our new added features. The prediction accuracy is very promising.

1. Introduction

There are many types of social networks that you can use through internet. The types are according to the functionality each network can provides [1]. One of these social networks is collaborative social network. Stack Overflow (http://stackoverflow.com), Quora, Yahoo!-answers and Never are examples of this type. In stack Overflow community the question answering process is conducting as follows. The user can choose a category to post a question to. After posting the question the asker waits specific amount of time in order to receive the answers from the expert users. Expert users are the users whose have a great knowledge in this category or sub category of the question field. If the question does not receive any answer, the asker can set a bounty to it. A bounty is a special reputation award given to answers. This feature was designed to motivate answerers, and help questions get the answers they deserve. Bounty awards are funded by the personal reputation of the users who offer them. Reputation is a rough measurement of how much the community trusts the answer author; it is earned if the answer author convinced other users that his answer is the best and this answerer knows what he is talking about. While bounty maker do not need to be the owner of a question to start a bounty on that question, only one bounty can be active on a question at once, and each user can only have up to three active bounties at once. Users must have specific reputation score to start a bounty, and at least as much reputation as the bounty amount. The bounty award will be subtracted from your reputation when the bounty is started, not when it is awarded.

If the question receives many answers, the users can give an up and down votes to both the question and answers in its answer thread. The most reputation points score is gained when the answer is up voted, it received a bounty, or it is selected as best answer .Also users can add comments to question and answers. Moreover users can set post as a favorite post. There are other activities but to be able to use them it depends on a user privilege under the community. User privilege depends on user reputation score. As an example of these privilege is to mark question or an answer as an offensive post.

[*] Corresponding author.
E-mail address: dallol_elalfy47@hotmail.com (D. Elalfy).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2017.06.002
1110-8665/© 2018 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.

Question's author can then select the most preferable or satisfied answer. This answer is called the best or the accepted answer. Unfortunately, this mechanism in question answering portals may lead to a lot of issues. Such as some askers cannot be able to choose the accepted answer.

For their question. The consequences are a lot of questions left as "not-answered" [2]. Even if the question is answered, there exist some probability that the answerer is not an expert in such category. And just answer it because he faced with that question when he/she opens the question answering portal. As a result the community lose one of its targets which is sharing knowledge because of the low quality answers that can be exists.

Moreover, the answerer that gives the low quality answer is capable of give a high quality one if that user was faced by the right question that he/she is expert in.

In this paper, we focus on the problem of exchanging and sharing of the knowledge in stack overflow community. And how to ease knowledge exchange by saving asker wasted time and effort that user exerts to find a satisfied answer. By predicting the best answer to the user, which considered the main problem in question answering portals.

The work in this paper is organized as follows. In Section 2, related work and a depth analysis to different knowledge exchange approaches will be introduced. Section 3, a study for the model in Ref. [3] which is used to give a local reputation score to answers and our proposed hybrid model for predicting best answer. Section 4, presents the experiment results and discussion. In Section 5 the conclusion is presented.

2. Related work

In order to improve the mechanism in which question answer portals work under, we need to focus on enhancing the question answer routing approaches and finding best answer techniques... There are a lot of efforts done by researchers in this field to overcome these problems; we are categorizing these studies to four categories: recommend right experts to a specific question, predicting the best answer, finding group of collaborative experts, and direct questions to an expert. All of these are solutions in order to improve the user satisfaction rate by giving high quality answers to him. Also to minimize the loss of time that is as a result of waiting for the right expert to answer the question.

2.1. Recommend experts to the current question

In Ref. [4] they tend to find the right expert to answer specific question under certain category. They proposed a hybrid model to find experts using user reputation, user authority, and user subject relevance. In evaluating their model they used Yahoo! Answer platform in Taiwan. Also called Yahoo! Knowledge plus. They assign different priority to terms according to their place like if the word is in answer post, question post, or in question title. One of the main issues in their technique that they do not consider the quality of posts posted by the expert. Since HITS take only the number of posts as an indication to the authority of that user.

In Ref. [5] their aim is to recommend an appropriate users to answer a specific type of. They split questions into two types an authority and affinity questions. They also recommend

A social network that is suitable to answer that type of question. Authors created a website that would allow user's from different social networks to ask questions in order to gain knowledge or social interaction. Social networks used are Facebook, Myspace, and Twitter. The authority question is the question that seeks information, the affinity question is the question that seeks social interaction or opinion. They build Expertise Estimation Algorithm to determine the objectivity levels of questioners and responders. This objectivity level is used later in determine the affinity and authority users. They does not consider the relationship between answer and question so that the answer might be off-topic.

In Ref. [6] Enterprise Social Network (ESN) service can help employees to collaborate and communicate effectively with colleagues, with customers and with suppliers. In this paper authors propose a model to better support question answering process in ESN, by using a graph analysis approach. Based on the questioner's initial input list of potential answerers, it can extract a shared-interest group of people, whose interest is similar to the initial list of potential answerers, and sort the group of people according to a score of interest distance, and then recommend them to the questioner. To evaluate its applicability, the method is implemented in KDWeibo the most popular ESN platform in China. The algorithms include three key concepts: Interest Distance, Aggregate Specialization Graph (ASG) and Specialization Sub Graph (SSG). One of the drawbacks exist in their model is that the method will not work if the user is not providing any initial list of potential answerers or the quality of the initial list is too low.

The Authors in Ref. [7] introduced a probabilistic framework to predict best answerers for questions. By tracking answerers' history, interests of answerers are modeled with the mixture of the language model and the Latent Dirichlet Allocation Model LDA. They also used both user authority and activity in predication. They utilizes two models to define user's interests. They calculated the likelihood probability based on user profile in order to predict the expert. Also they model the prior information of user which is made up of two parts the authority and user activity. They conducted their experiment using Isak CAQ in china. As a feedback to their work they need to further investigate their work in large scale dataset. And to use more accurate and different user activity and authority models.

In Ref. [8] Authors use a mechanism to filter online social streams received as well as enable them to interact with most similar users by personalizing the process of information distribution. Their framework identifies the most appropriate users to receive specific post by calculating the similarity between posts of the target user and the others. The platform used in research is stack overflow. Similarity is calculated based on user's social activity. User's social activity is an integration of both user interest in posts published and social activities of that user. Each user is represented by two vectors in vector space model. First vector is social pattern vector which contains influence attributes and user's distribution. The second vector contains bag of words as a post content's vector. Term frequency and inverse document frequency is used to weight terms of each vector. Then an aggregated linear model is applied to combine the calculated cosine similarity in two vectors.

2.2. Finding the best answer

In Ref. [9] they focus on finding best answer in massive online open courses in which users enroll in courses and to further understand it they can ask and answer question in the course forum. The experiment is conducted on openHPI MOOC platform. The users used machine learning through train four classifiers. They are bagging, naive Bayes, MultiPerceptron, and Random Forest using user features, thread features and content features. They used as a historical data the questions that has at least two answers. The training is performed on the answers of 416 questions.

Ref. [10] is a survey that the researchers found that there is a high correlation between posting a high quality question and getting a high quality answer. So they studied the features that most important in question to be found in order to get high quality answers. These question related features are tags and terms, length of the question, presence of an example that may help users to more understand the question. Asker related feature such as an asker reputation is also mentioned as a good indicator to the quality of the question post.

In Ref. [11] they predict best answer using Yahoo! Answers platform historical data. The experiment is conducted on small number of questions with at least five answers they put 13 criteria and asked five Amazon Mechanical Turk workers to rate the answers based on those features. To be sure that they use high correlated features, they match the rate given by worker with the actual assessment of the asker. The MTurk rating is from 1 to 5. They furthered their investigation by applying logistic regression classifier to their thirteen features. They found that their features are highly correlated. So they decided to use different features, the new features are extracted from answer and question post. Their data set consists from 116 questions and 575 answers.

In Ref. [12] the main aim of their research is to get the factors that affect selecting the best answer. They perform their investigation on 250 stack Exchange community which considered a large scale analysis. They found that users do not evaluate the answers by any criteria to select it as a best answer. They simply depend on the cognitive heuristic such as the space of this answer or the order of it in the answer thread. As the number of answers increases the users rely more in voting on the cognitive heuristic and that lead to less reliable evaluation by users.

In Ref. [13] they investigate the problem of low quality answers in question answering communities. They found that most of the work nowadays in this area focus on answer features to predict the best answer, they ignore the question type as a feature that decides the attributes that should be exist to determine the best answer to an existing question. They divided their work into two parts, first part is analysis to question type to find the most suitable answer features to train the model. Second part is that they train quality classifier based on question type and aggregate the overall answer quality score. Moreover they propose novel quality features to predict best answer. Their experiment is conducted on Yahoo! Answers platform with large data set approximately 50 thousand question answer pairs. They found that their hierarchy classifier can predict low quality answers more than high quality ones.

In Ref. [14] they propose question answering model to collaborative learning. Their system is for Indonesian high school as a part of E-learning process. They used Wikipedia as a free, web-based, multilingual encyclopedia. After posting a question and receiving answers and votes by the students in the system, the answers is evaluated by comparing it with Wikipedia database. And similarity percentage is provided also a link as a source of more information about that answer. Then all previous question and answers are entered into knowledge base to be a reference to new students or students that may will face the same problem in studying.

In Ref. [15] this paper the researchers trying to predict if the answer will be selected as the best answer by the asker or not. They conduct their research on stack Overflow site. They designed features to train a supervised classifier. Then they investigate the most influential features in prediction results. They used only answer's content features. The features are divided into contextual features which are the features that address the relation between the answers under the same thread, content features which are the statistical features of the answer itself, and finally the question-Answer relationship features such as the time lag between the posting time of question and the time of the response. They found that contextual features are the most important and influential ones in prediction results. They used large dataset consists of 196,145 answers. The classification process is accomplished using random forest classifier with two fold cross validation.

In Ref. [1] they conduct large experiment of Yahoo! Answers platform. In order to recommending top high quality answers based on reputation score function. Also they used answer content features and combine both content based method and reputation score method. They collected approximately 130 thousands dataset from four different categories in Yahoo! Answer platform. They found that the proposed non content method beats the benchmark link analysis method like HITS method in finding the user authority in this network. They performed an analysis and found that the distribution of the network's activities follows the power law distribution. Moreover they found that reputation non content method outperformed the content one.

2.3. Finding collaborative experts

In Ref. [16] this paper authors trying to find group of experts to collaborate to give an answer to a specific question. The idea of collaboration is to give the question thread a high lasting value. They used stack Overflow platform to conduct their experiments. They finds that question answer process is a collaborative effort that require input from different users. They introduce a user to user compatibility concept. And then present a mechanism to model this concept in question answer communities. The model is build using users compatibility, topical expertise since users must be expert in the same topic or category of question, and availability of users in specific time to collaborate to come up with only one high quality answer.

2.4. Route questions to a specific expert

Ref. [17], this paper the first reporting large scale analysis of answerer behavior on session level. The purpose of this study is to route new questions to experts on the topic of the question. And to reuse large scale data to satisfy asker needs. They answered many questions one of those is when the user tend to answer the question. And how the answerer choose the question to answer for. The answers to these questions help in developing a more reliable question recommendation model. Yahoo! Answer platform is used in.

This large scale analysis. They found that user participate in maximum two communities or categories. And that users chose to answer the questions that face them when entering the platform. They applied the analysis on all users, they do not consider super active users.

3. The proposed hybrid best answer prediction model

Our proposed model is consisting of two modules the first module is to predict the best answers using content features. And the second one is predicting the best answer using non-content features. After that we combine them in one hybrid model to get the best prediction result.

3.1. Content model

The model is an enhancement to the work in Ref. [15]. The original model is considering answer content features only. The model in Ref. [15] consisting from three parts: the features that deals with the content of the answer alone which are answer content features. Features that considering the relation between a question and its answers which are question-Answer features. And features that measure the degree of similarity between the answer and others answers under the same question which called context features. Our previous model in Ref. [3] added novel features under answer content features. The novel features are added as an enhancement and they are chosen based on the results of the original model in Ref. [15] (see Fig. 1).

The proposed system described in Fig. 2. It consists of three modules. The first module is the preprocessing module, in this stage the question and its answers are preprocessed to extract the tokens that are more effective in feature extraction phase. The second module is feature extraction, in this phase the novel features are added. The added features are added under answer content features only. The last phase is the classification, in which we chose to train more than one supervised classifier to get the most accurate results.

3.1.1. Content model experiment

As in our work in Ref. [3] we choose stack overflow as a platform to conduct our research on. We collected the ground truth data under an Academic category, which is a stack overflow category. Our data is classified to two classes' best answer and not-best answer class. Our data size is 18,000 answers. We have evaluated the previous model using three different classifiers, Random forest RF, logistic regression, and Naïve Bayes. We trained Random forest classifier with 10 fold cross validation. We applied and used 22 answer content features only. The features described in Table 1.

3.2. Non-Content model

The model is an enhancement to the work in Ref. [1]. The main difference is that the work in Ref. [1] is conducted on Yahoo-answer platform. Part of their work is to predict the best answer using non content feature which is reputation score of the user that answered the question. We applied the reputation score function here in stack overflow portal and notice it's affection in the best answer prediction accuracy.

3.2.1. Reputation based non-content model

Reputation based non-content model is calculated by the multiplication of two scores the first one is user confidence level score and the second one is user expertise level score. Fig. 3 shows the structure of the non-content model as in Ref. [1].

3.2.2. User's participation level

This is the function that computes the participation level for answerers and to measure their activeness. The participation function f(u_{ij}) is a sigmoid function which provides a high reward to the users with high number of answers and less reward to the users with low number of answers. All answers are under certain category. Moreover the reward is increased lesly when the number of answers exceeds a certain threshold.

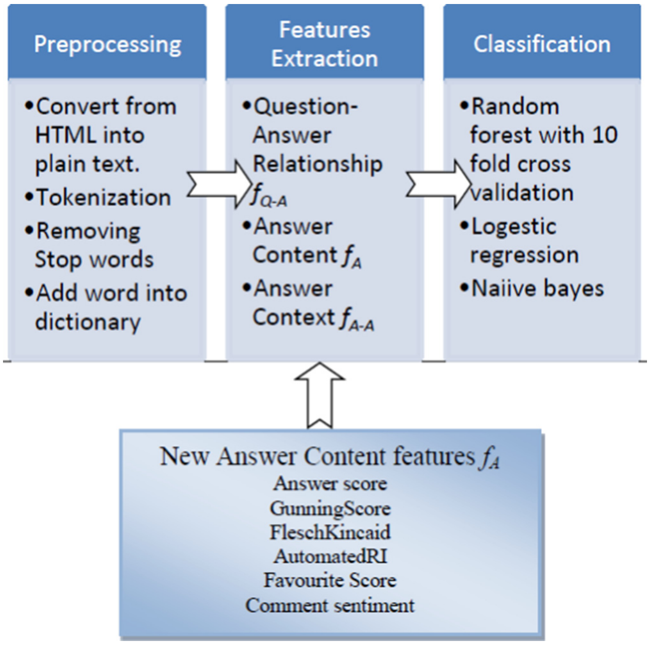
Preprocessing
•Convert from HTML into plain text.
•Tokenization
•Removing Stop words
•Add word into dictionary
Features Extraction
•Question-Answer Relationship $f_{Q\text{-}A}$
•Answer Content f_A
•Answer Context $f_{A\text{-}A}$
Classification
•Random forest with 10 fold cross validation
•Logestic regression
•Naiive bayes
New Answer Content features f_A
Answer score
GunningScore
FleschKincaid
AutomatedRI
Favourite Score
Comment sentiment

Fig. 2. Phases of content based model and new added features.

Table 1
Features designed under each type and their description.

Type	Symbol
f_A	ave_comment, var_comment comment_num URL_tag, pic, code ans_len readability
$f_{Q\text{-}A}$	QA-sim timeSlot
$f_{A\text{-}A}$	ave_ans_sim min_ans_sim max_ans_sim competitor_num ans_index

$$f(u_{ij}) = \frac{1}{1 + e^{\left(-\frac{x_{ij}-\mu}{\sigma}\right)}} \tag{1}$$

where x_{ij} is the total number of answers that user u_{ij} provides in category C_i and μ is the threshold value that determined based on the answer distribution pattern of the users. μ is used to reward

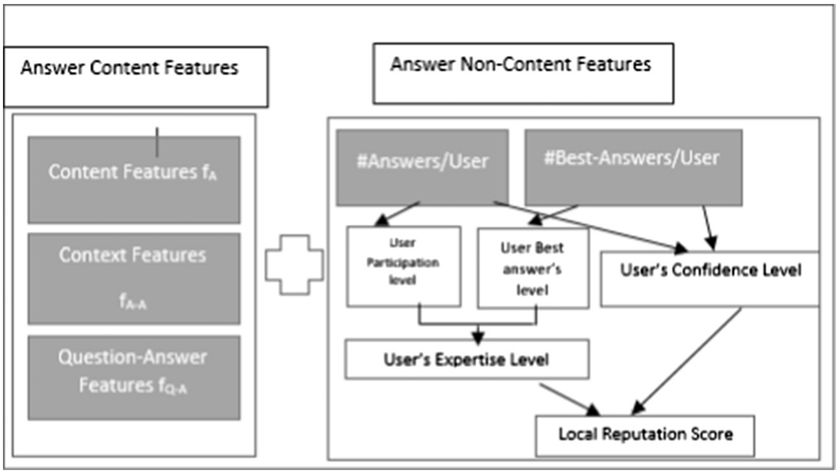

Fig. 1. (Hybrid model 1) Content and non-content model and their features description.

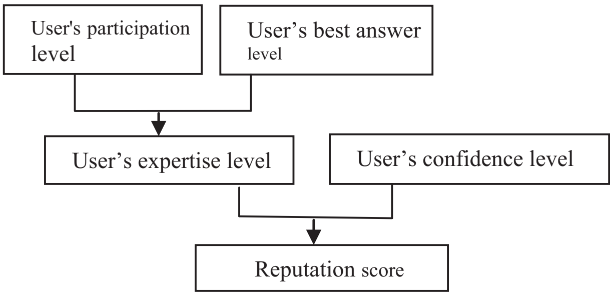

Fig. 3. User's reputation score.

most active users. Some highly active users provide more answers than other highly active users. To further reward those people the parameter σ is used which is the variation in the number of answers. It is calculated as

$$\sigma = \sqrt{\frac{(x_{ij} - \bar{x})^2}{t}} \tag{2}$$

where $\bar{x}$ is the average number of answers in the category i. and t is the total number of unique answer authors in the category i.

3.2.3. User's best answer level

This function is responsible for measuring the degree of expertise of the user under certain category. As in Ref. [1] they found that most of users have little number of answers selected as best answer and the minority have large number of best answers. So the best answer score $g(u_{ij})$ can be driven by applying the sigmoid function as in the participation function.

$$g(u_{ij}) = \frac{1}{1 + e^{\left(-\frac{y_{ij} - \mu_b}{\sigma_b}\right)}} \tag{3}$$

where y_{ij} is the total number of best answers that user u_{ij} provides in category C_i and μ_b is the threshold value that determined based on the best answer distribution pattern of the users. The parameter σ_b is used which the variation in the number of best answers is given by users in the category I. It is calculated as in Eq. (2).

3.2.4. User's expertise level

User's expertise level is influenced by the user's activeness and degree of participation in answering the questions. It's the average of both scores.

$$expe(u_{ij}) = \frac{f(u_{ij}) + g(u_{ij})}{2} \tag{4}$$

3.2.5. Confidence level

Confidence level $con(u_{ij})$ it's the ratio between the number of best answers $_{yij}$ to the total number of answers x_{ij} in category c_i by user $u_{ij.}$

$$con(u_{ij}) = \begin{cases} y_{ij}/x_{ij}, & if\ |y_{ij} > 0 \\ 0.0001, & if\ |y_{ij} = 0 \end{cases} \tag{5}$$

The confidence level function ensures that users who have high number of best answers have high score than others. To differentiate between new users who have small numbers of answers but most of them are best answers and the users who have large number of answers and the minority of them are best answers.

3.2.6. Reputation score

The reputation score of the user in a specific category is determined by both the confidence level $con(u_{ij})$ and expertise level $expe(u_{ij})$.

$$R(u_{ij}) = con(u_{ij}) \times expe(u_{ij}) \tag{6}$$

The value of R(uij) should be between 0 and 1. Zero means that the user uij is not an expert and one means that the user is an expert with a high reputation.

4. Experiments

4.1. Dataset description

In this research we used the same data set in the previous model in Ref. [3]. The ground truth data is collected from stack overflow portal and academic category. We separate our data into 4 windows, each window has different size. Window sizes are 8000, 11,000, 13,000 and 18,000 answer posts. The data is labeled with not-best answer and best answer class. We extracted the answerer information such as number of best answers answered by user in that category.

4.2. System description

In this paper, we have dealt with the problem of prediction if the given answer is going to be selected as the best answer or not based on the non-content features of this answer. We conducted our proposed model based on the work presented in Ref. [1].

We divided our work into three parts:

1- Content model as in our previous work in Ref. [3].
2- Non-content model which is a combination from user expertise level score and user confidence level score.
3- Merging content and non-content models with each other's.

4.3. Results and discussion

We used precision, recall and accuracy to measure the performance of our experiments. Precision measurement which is the proportion of the true positives t_{ps} versus all the positive results (true positive *tps* and false positive *fps*). True positive means the classifier correctly classified the case as best answer class. False positive means the classifier incorrectly classified the case as best answer.

$$Precision = \frac{t_{ps}}{t_{ps} + f_{ps}} \tag{7}$$

Recall measure which is the fraction of the true positives to all positive classes (true positives t_{ps} and false negative f_{ns}). False negative means the classifier incorrectly classified the cases to be in not best answer class.

$$Recall = \frac{t_{ps}}{t_{ps} + f_{ns}} \tag{8}$$

While accuracy measure is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$Accuracy = \frac{t_{ps} + f_{ns}}{t_{ps} + f_{ps} + f_{ns} + t_{ns}} \tag{9}$$

4.3.1. Results of using content features only

We chose to train Random Forest classifier, Logistic Regression and Naive Bayes using only content features. Random forest classifier with 200 trees and 10 fold cross validation while considering 5 random features. The random features are chosen by Weka framework.

In Table 2 we divided our ground truth data into parts each part called window size. We train the classifiers in stages each stage contains different window size. In order to inspect the impact of increasing the data set to the classifier accuracy. In window size 8000 answer set. The best prediction accuracy is 76.87% using logistic regression classifier. Then random forest which is 75.98% and naïve Bayes is the latest one by 73.98%.

In second window size which is 11,000 answers. As you can see the prediction accuracy decreases in both logistic regression and naïve Bayes. While it is increases by 5% in Random forest classifier which is a good indicator that random forest classifier is fit for our data set. Because it is a powerful and efficient classifier when dealing with large data set.

In window size 13,000, it is obvious that by increasing the window size, the behavior of the classifier does not change. So both logistic regression and naïve Bayes still decreases in the accuracy. They are 74.05% and 66.13% respectively. And random forest increases by 4% to become 84.12%.

The same behavior by classifiers in window size 18,000 except logistic regression is increased but slightly. And still random forest classifier beats the other. Its accuracy increased to be 88.36%. The most accurate prediction accuracy when using random forest classifier. In general in random forest as increasing window size the prediction accuracy increases.

Fig. 4 shows the comparison between the classifiers behavior with respect to different window sizes in x-axis and prediction accuracy in y-axis. It's clearly stated that random forest classifier beats others.

4.3.2. Results of using local reputation score only

The same analysis was applied to the data set using non-content feature only which is our proposed reputation score model RNC. Here the classifiers are trained using only the reputation score of the answer author and the predicition class. Table 3 shows the prediction accuracy results using different window sizes.

As in Table 3 you can see that the behavior of the three classifiers is almost the same random forest, naïve Bayes, and logistic regression prediction accuracy is decreased when the window size is increased. In window size 8000 answer set. The best prediction accuracy is 74.148% using random forest classifier. Then both logistic regression and naïve Bayes with 73.635%.

In second window size which is 11,000 answers. You can observe that the prediction accuracy decreased in all classifiers.

In window size 13,000, it is obvious that by increasing the window size, the behavior of the classifier does not change, it still suffering from decreasing in the prediction accuracy.

In window size 18,000 all classifiers results started to increase but still report results less that the first ones.

Fig. 5 shows the comparison between our local reputation classifiers behavior with respect to different window sizes in x-axis and prediction accuracy in y-axis. The classifiers report less accuracy as we increase the window sizes.

4.3.3. Hybrid model 1(using content and non-content features)

In this subsection, we investigated the affection of merging both previous content features and non-content features (local reputation). The prediction accuracy results are listed in Table 4.

Table 2
Best answer prediction accuracy in content classifier.

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.735	0.76	75.98%	8000
	0.808	0.808	80.84%	11,000
	0.842	0.841	84.12%	13,000
	0.884	0.884	88.36%	18,000
Logistic regression	0.747	0.769	76.87%	8000
	0.719	0.721	72.09%	11,000
	0.74	0.741	74.05%	13,000
	0.776	0.78	77.95%	18,000
Naïve Bayes	0.718	0.74	73.98%	8000
	0.686	0.683	68.33%	11,000
	0.7	0.661	66.13%	13,000
	0.738	0.663	66.82%	18,000

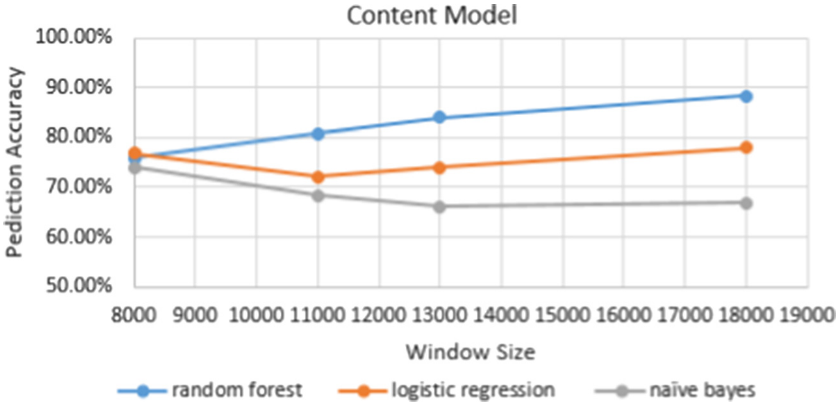

Fig. 4. Content classifiers results.

We found that adding non content feature to the content classifier does not affect the prediction accuracy positively. And does not decrease it a lot, the decrease in percentage is about 0.05 which is a non-significant change.

Fig. 6 shows the comparison between classifiers using both content and local reputation non content features under title hybrid model 1. The comparison with respect to different window sizes in x-axis and prediction accuracy in y-axis. Classifiers report almost the same behavior and accuracy as in using content features only. Still random forest classifier is the best and its accuracy is 88.34% occurred using window size 18,000 (see Figs. 7–10).

4.3.4. Results of using stack over flow reputation score only (SR)

We decided to use and apply the stack overflow reputation score instead of our local reputation model's score in order to investigate their effect in prediction accuracy. The following table shows the prediction accuracy of the three classifiers using only the reputation score given by stack overflow community (see Table 5).

Surprisingly, we found that using stack overflow reputation score in prediction does not make the high improvement in accuracy. And our reputation model and stack overflow reputation model give almost the same accuracy in prediction. Moreover, our reputation score was slightly higher than the stack overflow one.

Table 3
Best answer prediction accuracy in RNC Model (local reputation score only).

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.711	0.741	74.148%	8000
	0.633	0.64	63.96%	11,000
	0.628	0.626	62.60%	13,000
	0.661	0.678	67.75%	18,000
Logistic regression	0.542	0.736	73.635%	8000
	0.607	0.615	61.52%	11,000
	0.427	0.58	58.03%	13,000
	0.412	0.642	64.16%	18,000
Naïve Bayes	0.542	0.736	73.635%	8000
	0.607	0.615	61.52%	11,000
	0.587	0.587	58.73%	13,000
	0.412	0.642	64.16%	18,000

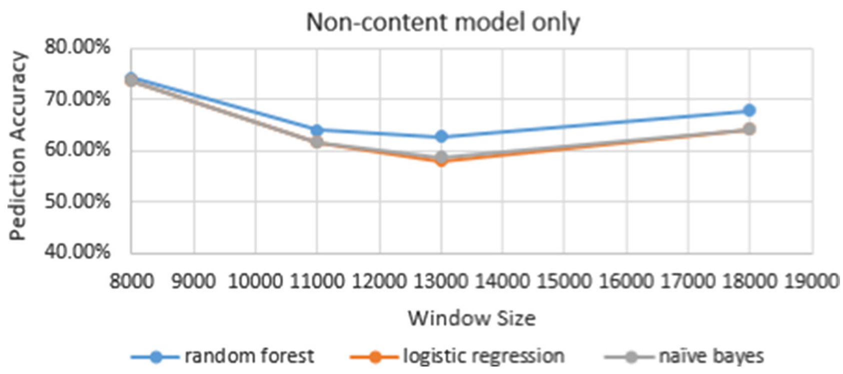

Fig. 5. Local reputation classifiers results.

Table 4
Best answer prediction accuracy after adding content and non-content features in classifier (hybrid model 1).

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.734	0.759	75.93%	8000
	0.804	0.803	80.32%	11,000
	0.159	0.841	83.96%	13,000
	0.884	0.883	88.34%	18,000
Logistic regression	0.747	0.769	76.88%	8000
	0.719	0.721	72.14%	11,000
	0.265	0.736	73.61%	13,000
	0.776	0.78	77.98%	18,000
Naïve Bayes	0.502	0.74	74.034%	8000
	0.685	0.682	68.22%	11,000
	0.738	0.666	66.57%	13,000
	0.738	0.666	66.58%	18,000

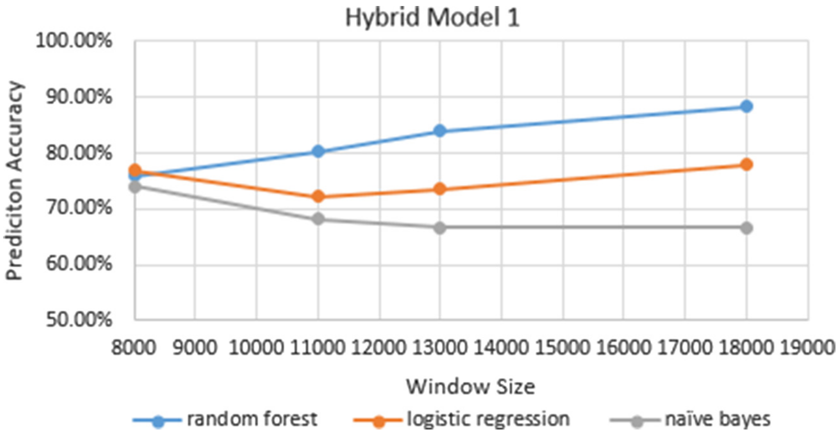

Fig. 6. Hybrid model 1 results.

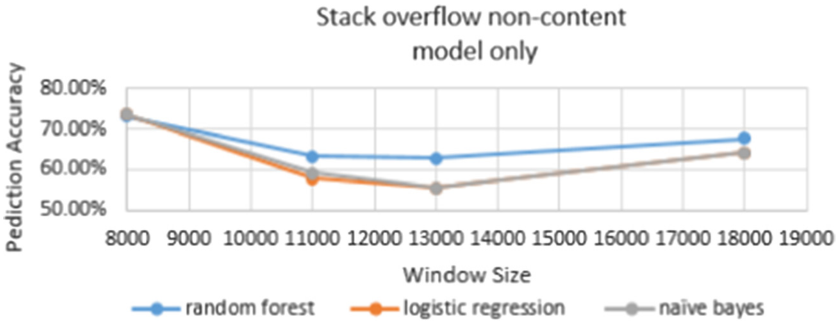

Fig. 7. Stack overflow reputation score results.

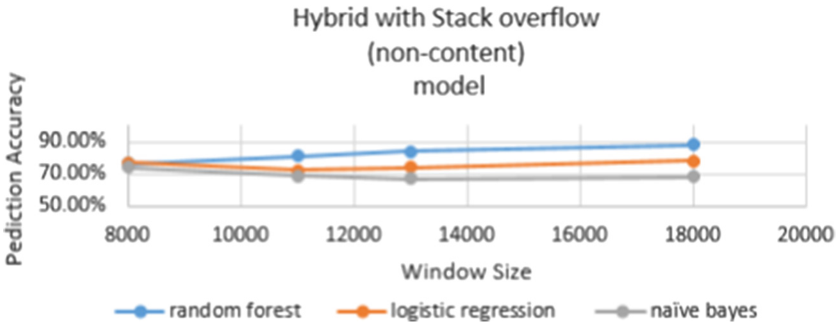

Fig. 8. Hybrid model 2 results.

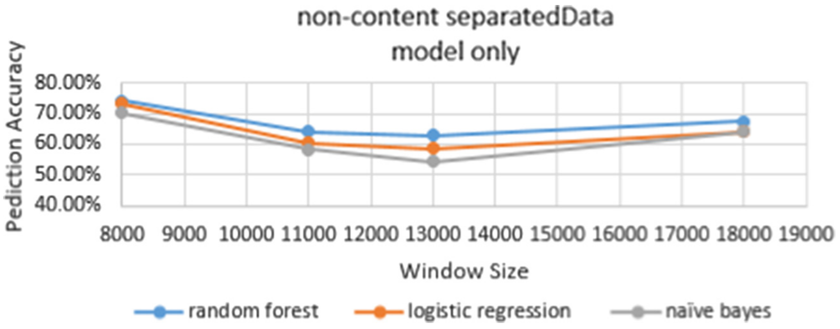

Fig. 9. Separated non-content results.

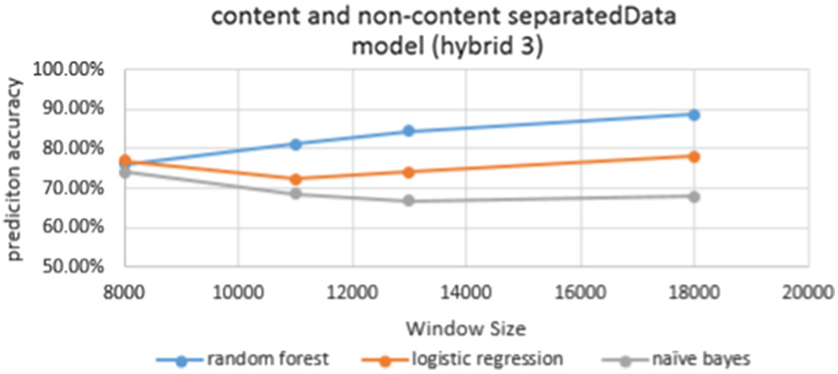

Fig. 10. Hybrid model 3 results.

4.3.5. Hybrid model 2

Table 6 captures the prediction accuracy of the classifiers using both stack overflow non content feature and content features.

We found that as we increase the window size the prediction accuracy increased and to compare hybrid model 1 HCR results with hybrid model 2 HSR, we found that in the first and the second window sizes the hybrid model 2 results is higher by 0.13%. And in the third and the fourth window sizes the opposite occurs. In general hybrid model 1 and hybrid model 2 are the same in prediction accuracy. And both of them is lower than the accuracy when we train the classifier using only the content features. The higher accuracy we got is when using the content features only it is 88.36%.

Using both content and stack overflow non-content reputation score in the classification process together is better than using stack overflow reputation score only. But still hybrid model 1 accuracy in prediction is higher than hybrid model 2.

4.3.6. Results of separated Non-content model

We expanded our analysis to investigate non content features separated. The non-content features used is number of participations to the answerer in this category of question and number of best answers of this user in this category. We reported the prediction results of classifiers in Table 7.

Table 5
Best answer prediction accuracy using non-content feature in classifier SR (Stack overflow reputation model only).

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.639	0.73	73.03%	8000
	0.624	0.633	63.25%	11,000
	0.628	0.627	62.73%	13,000
	0.659	0.677	67.69%	18,000
Logistic regression	0.542	0.736	73.64%	8000
	0.54	0.577	57.71%	11,000
	0.56	0.555	55.53%	13,000
	0.412	0.642	64.16%	18,000
Naïve Bayes	0.542	0.736	73.64%	8000
	0.572	0.592	59.15%	11,000
	0.566	0.553	55.32%	13,000
	0.412	0.642	64.16%	18,000

Table 6
Best answer prediction accuracy after adding content and stack overflow non-content feature in classifier (hybrid model 2 HSR).

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.737	0.761	76.08%	8000
	0.808	0.808	80.76%	11,000
	0.84	0.839	83.88%	13,000
	0.882	0.881	88.13%	18,000
Logistic regression	0.746	0.769	76.86%	8000
	0.717	0.72	71.97%	11,000
	0.738	0.738	73.77%	13,000
	0.778	0.781	78.13%	18,000
Naïve Bayes	0.719	0.741	74.05%	8000
	0.691	0.686	68.60%	11,000
	0.703	0.667	66.69%	13,000
	0.639	0.678	67.80%	18,000

Table 7
Best answer prediction accuracy using non-content separated features in classifier.

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.717	0.742	74.23%	8000
	0.633	0.64	64.01%	11,000
	0.631	0.63	62.95%	13,000
	0.661	0.67	67.47%	18,000
Logistic regression	0.633	0.64	73.17%	8000
	0.601	0.605	60.54%	11,000
	0.593	0.585	58.48%	13,000
	0.596	0.642	64.20%	18,000
Naïve Bayes	0.689	0.703	70.32%	8000
	0.559	0.538	58.26%	11,000
	0.56	0.544	54.42%	13,000
	0.412	0.642	64.16%	18,000

Table 8
Best answer prediction accuracy using both content and non-content separated features in classifier.

Classifiers	Prediction accuracy			Window size (#answers)
	Precision	Recall	Accuracy	
Random forest	0.738	0.762	76.22%	8000
	0.812	0.812	81.19%	11,000
	0.846	0.845	84.50%	13,000
	0.887	0.887	88.65%	18,000
Logistic regression	0.748	0.77	76.99%	8000
	0.722	0.724	72.39%	11,000
	0.741	0.741	74.11%	13,000
	0.779	0.782	78.12%	18,000
Naïve Bayes	0.719	0.741	74.09%	8000
	0.69	0.686	68.6%	11,000
	0.669	0.669	66.90%	13,000
	0.735	0.679	67.93%	18,000

Table 9
Comparison between the three non-content models in prediction accuracy.

Non-content classifiers	Prediction Accuracy			Window size (#answers)
	Random forest	Logistic regression	Naïve Bayes	
Our local reputation model 1(RP)	67.75%	64.16%	64.16%	18,000
Stack overflow reputation model 2(SR)	67.69%	64.16%	64.16%	
Separated non-content model 3(SP)	67.47%	64.20%	64.16%	

As you can see in Fig. 9, the three classifiers results is mostly the same. And the behavior is the same. The higher prediction result is when using smallest window size in all classifiers. And still random forest beats the others.

4.3.7. Hybrid model 3

Then we added both content with non-content separated features into the classification model to inspect the effect of the adding and the results are shown in Table 8. Non content separated features are number of participated answers by the user y in category x and number of best answers by user y in category x, respectively.

In Table 8, Hybrid model 3 is introduced. Contains both answer content features and non-content separated featues which is both number of participations to this answerer in the category and number of best answers to this answer in the same category of question. We found that as usual random forest classifier beats other classifiers in prediction accuracy. Finally, We got a promising results by hybrid model 3, since its prediction accuracy using 18,000 window size as well as random forest classifier is 88.65% (see Fig. 10).

4.3.8. Comparing results

In Table 9 we Compares the three non-content models in prediction accuracy**.**

In Table 9, we compared all non-content models in Tables 3, 5, and 7 consecutively with respect to the largest window size. We concluded that if you want to train a model to predict best answers based on non-content features only, you can use random forest classifier with our local reputation score feature instead of using either stack overflow reputation score or separated non-content features to get higher prediction accuracy.

In Table 10 we compared the three introduced hybrid models in Tables 4, 6, and 8, we found that the same trend is in all classifiers.

Table 10
Comparing three hybrid models in prediction accuracy.

Classifiers	Prediction accuracy			Window size (#answers)
	Random forest	Logistic regression	Naïve Bayes	
Hybrid model 1(HCR)	88.34%	77.98%	66.58%	18,000
Hybrid model 2(HSR)	88.13%	78.13%	67.80%	
Hybrid model 3(HSP)	88.65%	78.12%	67.93%	

Where as the window size increased so the prediction accuracy. And that the higher prediction accuracy is got using random forest which is a reasonably result, since random forest is a reliable and robust classifier that works better as the number of dataset increases. In hybrid model 1 in Table 2 which is the combination between content features and our local non- content reputation features. We found that random forest classifier accuracy using window size 18,000 is 88.34%. In hybrid model 2 in Table 4 which is the combination between content features and stack overflow non-content reputation feature. We found that random forest classifier accuracy using window size 18,000 is 88.13%. In hybrid model 3 in Table 8 which is the combination between content features and non-content separated features. We found that random forest classifier accuracy using window size 18,000 is 88.65%. We concluded our finding reporting that using both content and non-content separated features (hybrid model 3) in predicting the best answer is the best and giving us the higher accuracy than all others.

5. Conclusion

In this paper we worked on the question answering communities. And the issues that impede these communities to achieve their goals of sharing and exchanging to high quality knowledge. Then we discussed our previous model in predicating best answer in QACs and our findings using only content features. After that, we introduced an existing model using non-content features only, this model was used in different platform which is yahoo Answers. Also we compare the result of using the introduced non-content model with two different non-content models. We inspected the impact of using the non-content models alone. Finally, we introduced a different hybrid models which are combination of both our previous content and the non-content models. Moreover, we compared hybrid models results. We concluded the prediction results will increased by the merge process of hybrid model 3 with the content model and separated non-content features. Our experiments are conducted on stack overflow CQA platform .Also we found that using our introduced non-content model alone is better than using stack overflow non-content or using separated non-content model since it affects in increasing the predication accuracy positively.

References

[1] Chen Lin, Nayak Richi. Leveraging the network information for evaluating answer quality in a collaborative question answering portal. Soc Netw Anal Min 2012;2(3):197–215.

[2] Enterprise Social Network. In: Services Computing Conference (APSCC), Asia-Pacific: IEEE; 2012.

[3] Elalfy Dalia, Gad Walaa, Ismail Rasha. Predicting best answer in community questions based on content and sentiment analysis. In: 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE; 2015.

[4] Liu Duen-Ren et al. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites. Inf Process Manage (2013);49 (1):312–329.

[5] Zhan Justin, Jones Nadia M., Purnell Michael D. Top-K algorithm for recommendation in social networking kingdoms. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE; 2011.

[6] Ning Ke, Li Ning, Zhang Liang-Jie. Using graph analysis approach to support question & answer on enterprise social network. In: Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific. IEEE; 2012.

[7] Liu Mingrong, Liu Yicen, Yang Qing. Predicting best answerers for new questions in community question answering. Berlin Heidelberg: Web-Age Information Management. Springer; 2010. p. 127–38.

[8] Abeer ElKorany, ElBahnasy Khaled. Personalizing of content dissemination in online social networks. target 4.12 (2013).

[9] Jenders Maximilian, Krestel Ralf, Naumann Felix. Which answer is best?: predicting accepted answers in MOOC forums. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee; 2016.

[10] Baltadzhieva Antoaneta, Chrupała Grzegorz. Question quality in community question answering forums: a survey. ACM SIGKDD Explor Newsl 2015;17 (1):8–13.

[11] Shah Chirag, Pomerantz Jefferey. Evaluating and predicting answer quality in community QA. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM; 2010.

[12] Burghardt Keith et al. The Myopia of crowds: a study of collective evaluation on stack exchange. Robert H. Smith School Research Paper No. RHS 2736568 (2016).

[13] Toba Hapnes et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Inf Sci 2014;261:101–115.

[14] Arai Kohei, Handayani Anik Nur. Question answering system for an effective collaborative learning. IJACSA J (2012);3(1).

[15] Tian Qiongjie, Zhang Peng, Li Baoxin. Towards predicting the best answers in community-based question-answering services. ICWSM 2013.

[16] Chang Shuo, Pal Aditya. Routing questions for collaborative answering in community question answering. In: Proceedings of the 2013 IEEE/ACM International conference on advances in social networks analysis and mining. ACM; 2013.

[17] Liu Qiaoling, Agichtein Eugene. Modeling answerer behavior in collaborative question answering systems. Springer Berlin Heidelberg: Advances in information retrieval; 2011. p. 67–79.