World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:7, 2024

Identifying Quality Islamic Content in Community Question Answering Sites

Rabia Bibi, Muhammad Shahzad Faisal, Khalid Iqbal, Atif Inayat

***Abstract*—**Internet is growing rapidly and new community-based content is added by people every second. With this fast-growing community-based content, if a user requires answers of particular questions, then reviews are required from experts or community. However, it is difficult to get quality answers. The Muslim community all over the world is seeking help to get their questions and issues discussed to get answers. Online web portals of religious schools and community-based question answering sites are two big platforms to solve the issues of users. In the case of religious schools, there are experts and qualified religious scholars (mufti) who can give the expert opinion. However, the quality of community-based content cannot be guaranteed as it may not be an answer that satisfies the question of a user. Users on CQA sites may include spammers or individual criticizing the questioner instead of providing useful answers. In this paper, we research strategies to naturally distinguish the right content. As an experiment, we concentrate on Yahoo! Answers, and Quora, popular online QA sites, where questions are asked, answered, edited, and organized by a large community of users. We present the classification of data to categorize both relevant and irrelevant answers. Specifically, we demonstrate that the proposed framework can isolate quality answers from the rest with an exactness near that of people.

***Keywords*—**Community-based question and answering, evaluation and prediction of quality answer, answer classification, Islamic content, answer ranking.

I. Introduction

IN recent years, Community Question Answering (CQA) sites have emerged as a tremendous marketplace to satisfy the need of data. Estimating the number of questions that are answered on these websites is challenging, but it is likely that the volumes of questions with answers on such community answering websites far exceed that of library reference services [1]. Traditionally, library reference services were among the few reference services for such question answering. CQA websites make the questions and related responses submitted to the site accessible online and indexed via web indexes, thus enabling web users to find answers to previously asked questions in response to new inquiries.

Quora and Yahoo! Answers (YA) are two prominent examples of CQA websites; the use of these two sites has expanded significantly in recent years. The high usage level and vast amount of information on these platforms necessitate establishing criteria and rules for assessing the quality of explanations provided. Library reference services practice assessment to gauge the level of user satisfaction with the service [2], and this kind of assessment is no less important for the CQA websites. Unlike library services, Community Question Answering (CQA) websites lack an established set of expert guidelines and ethics. There is limited research on assessing the quality of answers provided on CQA websites. Although some research has been done to estimate user satisfaction [3], there is little work on identifying the components that can be used to evaluate the quality of a reply other than just being agreeable to the asker.

Liu et al. [3] took reference from the study of interactive question answering. The major difference, however, is that studying CQA websites involves real users. For example, in the use of questions in TREC QA, track submitted to the Finder system of FAQ, where replies are weighed by trained evaluators [4]. To ensure a large-scale assessment of structures similar to TREC track, it is appropriate to minimize the subjectivity in evaluation of these answers.

In any case, in CQA, the bias of importance appraisals is central. Furthermore, in evaluations of CQA, it is possible to gather specific evaluations from real clients – as opposed to prepared evaluators – potentially even the asker's own significance appraisals. Other than the benefits to the users of getting better options to assess the quality of these answers, having such metrics would also help the administration of CQA websites.

Most CQA websites have implemented ranking systems based on participation of the users on the website, where members earn score, or advance to higher levels based on criteria such as the number of questions answered and selection of answers by users as best answers. Having a mechanism for assessment of quality answers would benefit CQA websites by incorporating this factor into the creation and maintenance of reviewer's reputations. Ranking systems that implement these quality parameters also assist users who ask questions: askers can view user profiles and the history of responders to their questions, allowing them to assess the quality and ranking of previous answers. Recognizing the components that contribute to the quality of answers is crucial for web users in determining if a previously given answer is suitable.

In this research, we proposed an approach to measure the quality of Islamic answers on CQA websites and used it to predict which of the given responses would be selected by the asker as relevant. The specific issue of predicting the answer quality is outlined in Section III. In the following section, we introduce our classification method for measuring the quality of

Rabia Bibi, Ch. Muhammad Shahzad Faisal and Dr. Khalid Iqbal are with Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan (e-mails: rabia.inayat@yahoo.com, shahzad_faisal@ciit-attock.edu.pk, khalidiqbal@ciit-attock.edu.pk).

Atif Inayat is with Design and Development Department, APF, PAC Kamra (e-mail: aatif.inayat@yahoo.com).

answers. Section IV details how we conducted research to assess the quality of answers, and the analysis and results of the research are presented in Section V. An overview of some related works is given in the following section.

II. Background

CQA sites have been in existence for a considerable amount of time, primarily for product differentiation. These websites typically allow users to submit questions on various subjects, such as Yahoo! Answers, Quora, and Wiki Answers. While some sites have broad scopes, others restrict the topics in different ways. According to Shah et al. [5], CQA websites consist of three parts: a mechanism for users to submit questions, a platform for users to provide answers, and a community that engages in exchanging these questions and answers. The concept of question answering on online platforms dates back to the era of Bulletin Board and Usenet. In that sense, the concept of CQA websites is nothing new. However, dedicated CQA websites emerged on the web only in the past decade or so. The first CQA website, Korean Naver Knowledge iN, was launched in 2002, followed by the first English language CQA website, Yahoo! Answers, in 2005. Despite their relatively short history, there has been significant interest among research scholars in investigating various aspects of CQA platforms, including information-seeking behavior [6], resource selection [7], social comments [8], user incentives [9], comparisons with other QA services [10], and other information-related behaviors.

Indeed, certain destinations are subject-specific, for example, Stack Overflow, which focuses on questions about software coding, and Math Overflow, which restricts its extension to software coding research oriented questions of math. Some websites serve a particular client group, for example, HeadHunterIQ, which targets business recruiters. Additionally, some websites are designed to answer particular sorts of inquiries, for example, Homework Hub, which exclusively assists with homework-related questions. From the perspective of user satisfaction – with both the answer provide and the overall site experience – it would be beneficial for CQA websites to have a system for triaging queries. While the topic of questions would undoubtedly be a key factor in such system, other factors to consider include the quality of answers given on the website.

Performing this type of triage manually is relatively straightforward, albeit tedious, and is commonly performed by librarians [2], [11]. For instance, the QuestionPoint reference service, which manages global collaboration of library reference services, automatically conducts this type of triage by matching individual library profiles with inquiries [6]. However, the complexity of such triage cannot be compared to the unpredictability that a human triage can provide.

In the context of examining CQA, accessing participants directly, such as askers or answerers, can be challenging. To overcome this challenge, researchers have adopted various approaches that involve intermediaries for the askers. For instance, Kim et al. [6] examined the comments given by askers to determine the best reply. Others have employed third parties to stand in for askers: Harper et al. [7] used college undergraduates as intermediaries, while Liu et al. [8] utilized both subject specialists and paid laborers from Amazon Mechanical Turk. Each of these methodologies has obvious points of interest: the former leverages the asker's own words and assessment criteria, whereas the latter can gather more detailed evaluative data. The study described here follows the latter approach, utilizing intermediaries to gather evaluative information.

Various methods have been proposed to develop assessment criteria used to study of CQA. Liu et al. [3] utilized a rating system similar to that of the asker. Kim et al. [6] allowed assessment principles to emerge from observations made by askers during the selection of best answer. Proxies were used by Harper et al. [7] to assess the quality of answers according to criteria taken from the assessment of library reference services. Zhu et al. [12] proposed one of the most comprehensive sets of assessment rules for answers in CQA, which included six features derived from guiding principles for answerers on CQA sites and user comments. For our research, we will employ these six features to assess the quality of answers.

III. Problems of Predicting Answer Quality

In any type of data content, the quality of an answer can be subjective. Among various factors, assessing quality may rely on the relevance of the content, which may be challenging to quantify, especially in the context of CQA. Therefore, we, provide our own interpretation of value based on the information and objectives we have at hand.

An example is Yahoo! Answers, where questions are typically determined to be resolved if the community votes and chooses one of the replies as best answer, or the asker designate it as the best reply for their question. It is possible that multiple answers are of high quality, but only one of them is chosen as the best solution. Liu et al. [3] demonstrated that an asker selecting a reply as the best answer is indicative of satisfaction. However, it is important to note that the asker may choose not to select any reply as the best. Additionally, if the community votes in favor of a reply that the asker does not choose, it may indicate dissatisfaction. For our work, we will follow the principle of assessing asker satisfaction.

Further to the previous details, we define the problem of answer quality prediction, where our aim is to predict whether a given answer is deemed high quality by the asker. The objective of our research is to anticipate if the answer was selected by the asker as relevant or irrelevant. To achieve this, we will evaluate the quality of each answer based on several measures. Initially, we will utilize the six features as different aspects of answer quality. Next, we will use these features to categorize a response into either the "yes" class (selected as relevant), or the "no" class (not selected as relevant).

IV. Research Methodology

A. System Architecture

The system architecture of the proposed method comprises

of the following four components:
1. Data gathering
2. Data extraction
3. Data assessment by human experts
4. Classification

B. Data Gathering

Our dataset focuses on Islamic questions and comprises 100 questions and 585 Question Answer (QA) pairs. These questions cover various aspects related to the Basic Pillars of Islam, such as Tawhid, Salah, Siyām, Zakat, Hajj, as well as other Islamic laws like Jihad, Hijab, and Beard.

C. Extraction of Features

We extracted six features from our dataset of answers that relate to the quality of the content. These features are as follows:
1. Sentiments (classified as positive, negative or neutral)
2. Thumbs-up
3. Thumbs-down
4. Length of the answers (in words)
5. Number of answers of an answerer
6. Number of best answers of that answer

D. Data Assessment by Human Experts

Human editors labeled all the questions in our dataset for quality. These experts were independent from our team and provided unbiased evaluation. They analyzed answers deeply and categorized them into two categories: relevant (represented by 1) and irrelevant (represented by 0), based on the criteria of task orientation, relevance and solvedness. Using the results of human experts' assessments as labels, we then classified the answers based on the previously mentioned features. We compared the results of our classification with that provided by the human experts (described in part VI) to assess the effectiveness and accuracy of our method.

E. Classification Tools

Classification was performed in Weka using three classifiers:
1. Decision Tree
2. Random Forest
3. Adaboost M1

1.Decision Tree

A decision tree is a classification tool that uses a tree-like structure to make decisions and their possible outcomes. It also includes probable event results, costs of resources utilized, and the utility of decision. It is just one method to display an algorithm. Decision trees can be used as a model for decision problems under uncertainty. They help to visualize the possible options, the events that might occur, and the consequences as a combination of decisions and events. Probabilities are assigned to events, and weights are calculated for each outcome. The most important role of using a decision tree is to analyze the available options and reach the best decisions.

Decision tree remained the best classifier by resulting in the highest truly classified values and lowest wrongly classified values.

2.Random Forest Classifier

The Random Forest classifier is a combination of tree indicators such that each tree depends on the approximation of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as the number of trees in the forest becomes large. Using a random selection of features to split each node yields error rates that compare favorably with AdaBoost but are more robust to noise. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and allowing them to vote in favor of the most popular class. In order to build these ensembles, random vectors are often generated to control the growth of all the trees in the ensemble.

3.AdaboostM1

AdaBoostM1 is a method used to improve the performance of any learning algorithm. It is typically used to significantly decrease the errors of any learning algorithm that generates classifiers which need to perform better than random guessing. Theoretical results have shown that boosting has potential benefits, but the practical values of boosting can only be measured by testing this method on real problems.

V. Analysis and Results

Using previously mentioned six features, our experts analyzed the dataset and classified the answers as "relevant" or "irrelevant". We used this classification (relevant as "1" and irrelevant as "0") as a label and classified the dataset using the aforementioned classifiers.

It was observed after applying the classifiers that the three most important and effective features influencing the results are Length, Sentiments, and Thumbs-up (Figs. 1-3), with effectiveness in descending order.

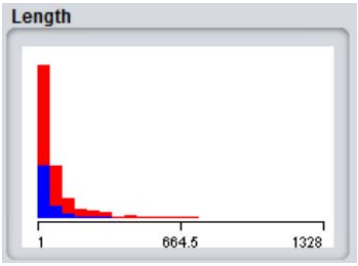

Fig. 1 Yahoo! Answers Length (Red = relevant, Blue = irrelevant, X-axis is the feature)

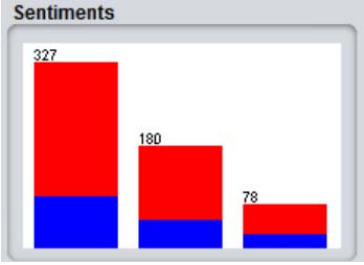

Fig. 2 Yahoo! Answers Sentiments

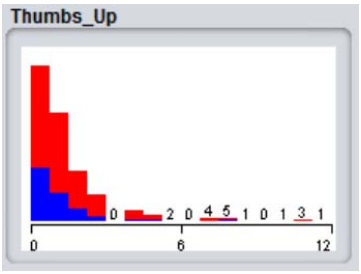

Fig. 1 Yahoo! Answers Thumbs-up

Decision Tree classification was successful 70.60% of the time in correctly classifying the relevant answers in the training set, where 413 out of 585 instances were correctly classified as show in Table I.

TABLE I
RESULTS OF DECISION TREE CLASSIFICATION

Decision Tree		
Result	Percentage	No. of Instances
Correctly classified	70.60%	413
Incorrectly classified	29.40%	172

Random Forest classification was successful 66.84% of the time in correctly classifying the relevant answers in the training set, where 391 out of 585 instances were correctly classified as shown in Table II.

TABLE II
RESULTS OF RANDOM FOREST CLASSIFICATION

Random Forest		
Result	Percentage	No. of Instances
Correctly classified	66.84%	391
Incorrectly classified	33.16%	194

AdaBoostM1 classification was successful 68.89% of the time in correctly classifying the relevant answers in the training set, where 403 out of 585 instances were correctly classified, and 182 were incorrectly classified as shown in Table III below

TABLE III
RESULTS OF ADABOOSTM1 CLASSIFICATION

AdaBoostM1		
Result	Percentage	No. of Instances
Correctly classified	68.89%	403
Incorrectly classified	31.11%	182

As can be seen in Table IV, Decision Tree had the highest F-score, while Random Forest had the lowest F-score value. Although the precision of the Decision Tree was not the highest, the Recall value was at the highest, resulting in the highest F-score.

TABLE IV
OVERALL PRECISION, RECALL, AND F1 SCORE OF OUR SELECTED CLASSIFIERS

Classifier	Precision	Recall	F-score
Decision Tree	0.717	0.976	0.827
AdaBoostM1	0.714	0.945	0.813
Random Forest	0.741	0.826	0.781

Table V displays the results of True Positive Rate, False Positive Rate, and ROC for all the three classifiers used. Random Forest exhibits the highest True Positive Rate, Decision Tree shows the highest False Positive Rate, and AdaBoost has the highest ROC Area. Therefore, a mixed trend is evident in these results.

TABLE III
TRUE POSITIVE RATE, FALSE POSITIVE RATE AND ROC AREA FOR THE TASK OF CLASSIFYING RELEVANT AND IRRELEVANT ANSWERS

Classifier	TP Rate	FP Rate	ROC Area
Decision Tree	0.976	0.982	0.569
AdaBoostM1	0.945	0.964	0.587
Random Forest	0.826	0.733	0.578

VI. CONCLUSION

Assessing the quality of content in CQA websites, especially in the context of religion, poses significant challenges. Information retrieval itself is complex, and content evaluation adds another layer of complexity in CQA platforms. Therefore, novel approached are needed rather than relying on the traditional rules of relevance as described by Saracevic [13]. In our study, we utilized six features to gauge the content quality of Yahoo! Answers. Through human assessment based on task orientation, relevance, and solvedness, we categorized answers into relevant and irrelevant categories, establishing a gold standard for comparison with our models. While this approached allows us to evaluate and predict content quality, we also identified other aspects such as completeness, informativeness, and novelty. However, these features alone were not sufficient to ensure high-quality content. Our human experts lacked context about the askers or answerers, and they were unaware of the best answers. We recognize the importance of providing such contextual information in the evaluation content quality of CQA platforms.

We extracted six features from both the answers and the answerers. Through the model construction process, it was revealed that the information gathered by the content and answerer's profile, significantly influences content quality assessment.

Beyond the selection of best answers and content evaluation, there are crucial considerations to bear in mind for content quality assessment in CQA. As mentioned earlier, the diversity in question and answer categories on CQA services is vast, and a question may attract multiple answers. Since only one answer can be the selected as the best answer, constructing a classifier to evaluate answers based on limited information provided by features is extremely challenging. However, by leveraging appropriate features, we were able to identify the relevant content.

Content quality evaluation in CQA presents unique opportunities to consider context and social factors. For example, information disposition, and the answerers' profile provide datasets that are extremely helpful for predicting and evaluating answers. Multiple CQA sites coexist, each with different mechanisms for asking, answering, or rating content, suggesting the presence of several other features of questions

and answers in CQA sites. In the future, further exploration of these features alongside our presented features will be essential for evaluating content quality in CQA comprehensively.

REFERENCES

[1] S. Kim, J. S. Oh, and S. Oh, "Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective," *Proceedings of the American Society for Information Science and Technology*, vol. 44, no. 1, pp. 1–15, Oct. 2008.

[2] T. Saracevic, "Evaluation of evaluation in information retrieval," pp. 138–146, Jan. 1995. (Online). Available: http://dl.acm.org/citation.cfm?id=215351. Accessed: Nov. 8, 2016.

[3] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, 2008. (Online). Available: http://www.mathcs.emory.edu/~eugene/papers/sigir2008-cqa-satisfaction.pdf. Accessed: Nov. 8, 2016.

[4] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence,* vol. 14, p. 1612, 1999

[5] E. M. Voorhees, "The TREC robust retrieval track," *ACM SIGIR Forum*, vol. 39, no. 1, p. 11, 2005. (Online). Available: http://trec.nist.gov/pubs/trec12/papers/ROBUST.OVERVIEW.pdf. Accessed: Nov. 8, 2016.

[6] D. N. Kresh, "Offering high quality reference service on the web: The collaborative digital reference service (CDRS)," Jun. 2000. (Online). Available: http://dl.acm.org/citation.cfm?id=865252. Accessed: Nov. 9, 2016.

[7] J. James, "The Global Census of Digital Reference," *5th Annual VRD Conference*, 2003

[8] Harper, F. Maxwell, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online Q&A sites," pp. 865–874, Jun. 2008. (Online). Available: http://dl.acm.org/citation.cfm?id=1357191. Accessed: Nov. 8, 2016.

[9] Q. Su, D. Pavlov, J.-H. Chow, and W. C. Baker, "Internet-scale collection of human-reviewed data," pp. 231–240, Aug. 2007. (Online). Available: http://dl.acm.org/citation.cfm?id=1242604. Accessed: Nov. 8, 2016.

[10] Z. Zhu, D. Bernhard, and I. Gurevych, "A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites," 2009. (Online). Available: http://tuprints.ulb.tu-darmstadt.de/1940/1/TR_dimension_model.pdf. Accessed: Nov. 8, 2016.

[11] C. Shah, U. C. Hill, J. S. Oh, and S. Oh, "Exploring characteristics and effects of user participation in online social Q&A sites," *First Monday*, vol. 13, no. 9, Sep. 2008. (Online). Available: http://www.firstmonday.dk/ojs/index.php/fm/article/view/2182. Accessed: Nov. 8, 2016.

[12] L. Rokach and O. Maimon, "Introduction to Decision Trees," *Data Mining with Decision Trees: theory and application,* p. 5, 2008.

[13] C. Shah, S. Oh, and J. S. Oh, "Research agenda for social Q&A," *Library & Information Science Research*, vol. 31, no. 4, pp. 205–209, Dec. 2009. (Online). Available: http://www.sciencedirect.com/science/article/pii/S074081880900098X. Accessed: Nov. 8, 2016.