



Assessing the quality of answers autonomously in community question–answering

Long T. Le¹ · Chirag Shah² · Erik Choi³

Received: 9 January 2017 / Revised: 28 June 2019 / Accepted: 29 July 2019 / Published online: 5 August 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Community question–answering (CQA) has become a popular method of online information seeking. Within these services, peers ask questions and create answers to those questions. For some time, content repositories created through CQA sites have widely supported general-purpose tasks; however, they can also be used as online digital libraries that satisfy specific needs related to education. Horizontal CQA services, such as Yahoo! Answers, and vertical CQA services, such as Brainly, aim to help students improve their learning process via Q&A exchanges. In addition, Stack Overflow—another vertical CQA—serves a similar purpose but specifically focuses on topics relevant to programmers. Receiving high-quality answer(s) to a posed CQA query is a critical factor to both user satisfaction and supported learning in these services. This process can be impeded when experts do not answer questions and/or askers do not have the knowledge and skills needed to evaluate the quality of the answers they receive. Such circumstances may cause learners to construct a faulty knowledge base by applying inaccurate information acquired from online sources. Though site moderators could alleviate this problem by surveying answer quality, their subjective assessments may cause evaluations to be inconsistent. Another potential solution lies in human assessors, though they may also be insufficient due to the large amount of content available on a CQA site. The following study addresses these issues by proposing a framework for automatically assessing answer quality. We accomplish this by integrating different groups of features—personal, community-based, textual, and contextual—to build a classification model and determine what constitutes answer quality. We collected more than 10 million educational answers posted by more than 3 million users on Brainly and 7.7 million answers on Stack Overflow to test this evaluation framework. The experiments conducted on these data sets show that the model using random forest achieves high accuracy in identifying high-quality answers. Findings also indicate that personal and community-based features have more prediction power in assessing answer quality. Additionally, other key metrics such as F1-score and area under ROC curve achieve high values with our approach. The work reported here can be useful in many other contexts that strive to provide automatic quality assessment in a digital repository.

Keywords Community question–answering (CQA) · Answer quality · Features · Education · Focused CQA

1 Introduction

The way people share and seek information has been undeniably changed by critical tools such as the Internet and the World Wide Web (WWW). Many online resources on the WWW serve as some of the largest digital libraries publicly available. As the number of new resources for communication and information technologies have rapidly increased over the past few decades [22], users have embraced such sources in their information seeking and sharing episodes. Examples of those sources include Wikis, forums, blogs, and community question–answering (CQA). Particularly, CQAs are a new means of information seeking in which users share information and knowledge in virtual environments.

✉ Long T. Le
longtle@rutgers.edu

Chirag Shah
chirags@rutgers.edu

Erik Choi
erikc@dropbox.com

¹ Department of Computer Science, Rutgers University, Piscataway, NJ, USA

² School of Communication and Information, Rutgers University, New Brunswick, NJ, USA

³ Dropbox, New York, NY, USA

According to Gazan [14], CQA is “exemplifying the Web 2.0 model of user-generated and user-rated content” (p.2302), creating a critical online repository and an engagement platform where users formulate their information need in natural language and voluntarily interact with each other through the asking and answering of questions. Beyond Q&A services, CQAs offer other means of interaction, such as commenting and voting, that encourage social information seeking and sharing. CQA popularity has exploded, and a correspondingly rich body of research has been conducted in order to understand the variety of content and user behaviors present within such platforms. According to Shah et al. [30], studies have been conducted on both user content and user behaviors. The former have focused on content type, quality, and formulation, while the latter attempted to understand the motivations for asking and answering a question on CQA.

Most initial CQA platforms—including AnswerBag, the first to appear in the USA—were developed to support general-purpose information seeking behaviors. These are referred to as horizontal CQA services. Later-developed “vertical” CQA sites pertained to more specific tasks, such as asking and answering a question to share educational information for learning in specific topics (e.g., Mathematics, Physics, etc.). Beyond looking for new materials, students can use the Internet to exchange knowledge to satisfy their educational needs. The advent of CQA has greatly assisted students in sharing knowledge in virtual environments. As CQA in education is an emerging field, educators hope that communication and information technologies may be able to improve learning capability and experience to improve students’ learning engagements in virtual environments.

To support CQA’s educational potential, we attempt to examine Brainly,¹ one of the largest CQA services specifically geared toward education. Brainly is an online social learning community for students and educators with millions of active users. It has approximately 80 million monthly unique visitors as of January 2017 and is available in 35 countries, including the USA, Poland, Russia, Turkey, Brazil, France, and Indonesia. Figure 1 shows the homepage of Brainly in the USA.

We also evaluate our framework through another popular CQA, Stack Overflow. Stack Overflow is a CQA site focusing on programming-related questions and is used widely by the programming community. Unlike general, horizontal CQA sites, all content should be related to the subject of programming. Stack Overflow’s data dump was released to the public.² Figure 2 shows its homepage.

CQA is a user-driven community where registered users voluntarily create all contents, including questions and answers. Therefore, content quality can easily vary, and it is

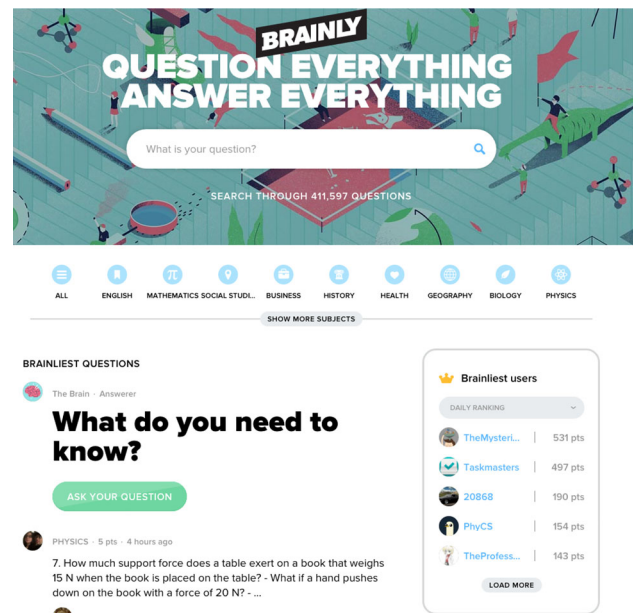


Fig. 1 Brainly’s homepage in the USA

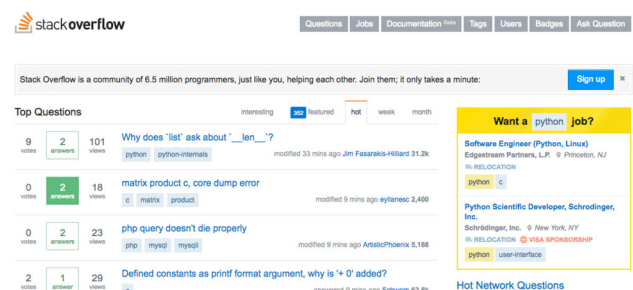


Fig. 2 Stack Overflow’s homepage

paramount that sites offer high-quality information to retain existing users and attract new users to support their online information seeking behaviors. The quality of information for educational purposes is even more critical, as educational CQA users may develop their knowledge base through interactions with contents within the platform. For example, it may be possible that students who use a CQA to ask questions about homework problems could be misled by incorrect answers without making proper judgments about answer quality. Struggling students, in particular, may suffer when faced with inaccurate information. Thus, quality assessment is critical for students to seek and share the right information for their learning on educational CQA sites. At the moment, traditional CQAs most likely depend on human judgments to evaluate content quality. Unfortunately, human assessors have many drawbacks, including subjective (and possibly biased) assessments, seeming difficulty in recruiting enough users who may be able to assess contents, and the time it could take for human assessors to go through the

¹ <http://brainly.com>.

² <https://archive.org/details/stackexchange>.

ever-increasing content in CQA sites. Here, we address these concerns with a new framework for assessing content quality. Our specific contributions are as follows.

- Conducting an empirical study. This is the first large-scale study to investigate the quality of answers in both an emerging CQA for education and one of the most popular subject-focused CQAs.
- Proposing a framework to assess the answers automatically. Our framework extracts different aspects of CQA content—such as personal features, community features, textual features, and contextual features—to build high-accuracy classifiers. Our work can achieve accuracy higher than 78.5% in all data sets. Our method also achieves high values on other key metrics such as F1 score and area under ROC curve.
- Examining the importance of different features and groups of features in assessing the quality of answers. The results show that personal features and community features are more important and have more predictive applications.

The rest of this article is organized as follows: Sect. 2 discusses the background and a few related works. The framework is described in Sect. 3. Section 4 presents the data sets used in our study. We present the results and discussion of our method in Sects. 5 and 6. Finally, the conclusion and future work are presented in Sect. 7.

2 Background and related works

2.1 Community question–answering (CQA)

Community question–answering (CQA) services have become popular places for online information seeking. Some popular CQAs, such as Yahoo! Answers and Stack Overflow, attract millions of users. CQA takes advantage of wisdom of the crowd, the idea that nobody knows everything, but everyone knows something [35]. Users can contribute to the community by asking questions, giving answers, and voting for posts. Humans moderate most activities.

A number of previous research works have investigated user interests and motivations for participating in CQA [27], [40]. Adamic et al. [1] studied the impact of CQA. In the work, the authors analyzed questions and clustered them based on the questions' contents. The results showed a diversity of user types in CQA. For example, some users can participate in a large number of topics, while many users are only interested in a narrow topical focus. Using some basic features such as response length and past answers given by the corresponding user, the work also examined the best answers. Shah et al. [29] compared CQA and virtual

reference to identify differences in users' expectations and perceptions. By understanding and identifying these behaviors, challenges, expectations, and perceptions within the context of CQA, we can better highlight potential strategies that more accurately match question askers with appropriate answerers. Le and Shah [19] developed a framework to detect the top contributors in their early stage by integrating different signals of users. These “rising star” users are crucial to CQA community health due to their high quantity of quality contributions.

2.2 Focused CQA

CQA has successfully supported Q&A that pertains to specific subjects such as programming, mathematics, and music. Stack Overflow, for instance, is a popular site that attracts millions of programmers. All questions in Stack Overflow must be related to programming. Within the site, users can demonstrate their expertise through knowledge sharing activities. In fact, many programmers are hired due to their high contribution in Stack Overflow [37]. Additionally, Le et al. [19] showed that there is a small number of top contributors in Stack Overflow and it is possible to detect these users in their early stages, while Yao et al. [44] studied the long-term impact of contents in the site and demonstrated that many answers with lasting significance could be considered as technical knowledge for many programmers.

2.3 Educational CQA

In recent years, online learning has collapsed time and space [9], which allows users to access information and resources for educational purposes any time and from anywhere. As online learning grows in popularity, a variety of new online information sources have emerged and are utilized in order to satisfy users' educational information needs. For example, empirical investigations have been conducted on social media (e.g., Facebook, Twitter, etc.) in order to understand the effectiveness of higher education [39]. Khan Academy has become a popular online educational video site that has more than 200 million viewers as well as approximately 45 million unique monthly visitors [25]. Additionally, even though most CQAs are mainly focused on seeking and sharing information that pertains to either general topics (e.g., Yahoo! Answers, WikiAnswers, etc.) and/or professional topics (e.g., Stack Overflow, etc.), new CQAs have emerged to help students participate in question–answering interactions concerning educational information for online learning. While some small-scale CQA tools were developed to support a group of university students [2,33], large educational

CQAs such as Chegg,³ Piazza,⁴ and Brainly have also introduced a new platform to support students' online information seeking via question–answering interactions in school subjects and topics. Brainly specializes in online learning for students (i.e., middle school, high school) through asking and answering activities in 16 main school subjects (e.g., English, Mathematics, Biology, Physics, etc.) [6].

2.4 Quality assessment in CQA

Since most contents in CQA are generated by users who actively seek and share information with other users, content quality is a critical factor to community success. It follows that assessing the quality of posts in CQA is a critical task in order to develop an information seeking environment where users receive reliable and helpful information, particularly in the educational realm. High-quality content is the best way to retain existing users and attract new members [23]. However, assessing the quality of posts in CQA is a difficult task due to the diversity of contents and users. A total post's standing may include the quality of both its question and answer(s). In our work, we focus on the quality of the answer.

The capability of detecting an untruthful contribution is also an important task [26,38]. Pelleg et al. [26] studied truthfulness in CQA sites. This research examined whether a user provides truthful information about themselves on CQA sites. They found that the askers generally provide accurate personal information, even when they post sensitive questions. Tan et al. [38] proposed a new method called CQAL to automatically predict the quality of a post in a social knowledge base. In a social knowledge base such as Wikipedia or Freebase, users can edit articles. These contributions might contain inaccurate information and reduce the user experience. Some signals that prove useful in determining false posts are user contribution history; the features of each triple including subject, predicate and object, and the user expertise. Examining answer quality can be divided into three tasks: (i) finding the best answer, (ii) ranking the answers, and (iii) measuring the quality of answers. For example, Shah and Pomerantz [31] looked for the best answers in Yahoo! Answers by using 13 different criteria. Ranking answers is a useful task when a question receives multiple answers. These works focus more on the similarity between an answer and a question [34]. Le et al. [21] showed that there are a number of users who want to participate in a CQA but fail to generate sufficient content. Suryanto et al. [36] utilized the expertise of both an asker and an answerer to rank answers. In this work, the authors recognized that different users hold expertise in different subjects, which factors into answer ranking. Recent work also showed the potential of ranking users with

graphs [15], but it is not yet clear how to rank answers based on users.

The most popular research inquiry focuses on regression-related problems, such as predicting how many answers a question will receive or how much community interest a post can elicit. Researchers are interested in predicting whether certain questions in CQA will be answered and how many answers a question will receive [12,41]. Such research used features such as asker history, question length, and question category to predict the answerability of a question. Shah et al. [32] studied why some questions remain unanswered in CQA. Particularly, their work explored why fact-based questions often fail to attract an answer. Momeni et al. [24] applied machine learning to judge the quality of comments in online communities, revealing that social context is a useful feature. Yao et al. [44] examined the long-term effect of the posts in Stack Overflow by developing a new scalable regression model. Dalip et al. [11] tried to reduce the number of features in collaborative content, but did not achieve a significant number of reductions. Furthermore, applying feature selection can solve the issue with many features, such as over-fitting.

Our framework follows the machine learning algorithm which includes two steps: training the model and apply the model. In the training step, the framework collects the label data and the list of features. In this study, the label is the quality of the answer. Then, we can build the model based on these label data to see the relationship between the features and the label. The built model is used to predict new data without the label. In this case, the label is the quality of answers. Extracting the right features is critical to the accuracy of the model. In a previous work, the textual features such as the length and the readability of the text are used widely [10,20,41]. The information of the users such as their activities also shows the efficacy in predicting the quality [11,43]. In this work, we also use egonet features of a node. A node's egonet is the subgraph created by the node and its neighbors. Egonet features include the size of egonet, the number of outgoing edges of egonet, and the number of neighbors of egonet. These features incorporate four social theories: Social Capital, Structural Hole, Balance, and Social Exchange [3]. The capacity of social connection in information dissemination was conducted in [18]. These theories capture connectivity of a user and his friends, control of information flow, transitivity among the users, and reciprocity among the users. Our work is the first work that uses these features in predicting the quality of answers. Another group of features is also useful in predicting the quality of the answer is contextual features. Contextual features record the context in which the answer is given. In our work, we examine carefully some contextual features such as the typing speed and the time that users spent on answering questions.

³ <https://www.chegg.com>.

⁴ <https://piazza.com/>.

Our work is focused on measuring the quality of answers by employing past question–answering interactions and current question and/or answering activities in order to automatically predict the quality of new answers. The framework incorporates different types of features, including personal features, community features, textual features, and contextual features. The early version of our work was conducted in educational Q&A [20].

3 Examining the quality of an answer

In order to reduce the workload involved in manually assessing answer quality, we developed a framework to automatically detect the quality of answers. This was a difficult task due to the complexity of CQA content. Here is the formal definition of our problem:

Formal definition:

Given:

- a set of users $U = \{u_1, u_1, \dots, u_n\}$
- a set of posts $P = Q \cup A$,
 Q is the set of questions $Q = \{q_1, q_2, \dots, q_{m1}\}$, and
 A is the set of answers $A = \{a_1, a_2, \dots, a_{m2}\}$
- a set of interactions $I = \{i_1, i_2, \dots, i_{m3}\}$ (such as giving thanks, up-votes, making friends)

Task For arbitrary answer $a \in A$, predict whether a will be *deleted* or *approved*?

Our framework follows a classification problem. In the first step, we collect users' history and information, the interactions in the community, and the characteristics of answers. In the second step, we build the classification model based on history. In the last step, we predict the quality of new answers based on our trained models.

3.1 Feature extraction

In order to classify answer quality, we build a list of features for each answer. Tables 1 and 2 list the features used in our study. Table 1 is used in Brainly and Table 2 is used in Stack Overflow. The slight differences are due to the characteristics of two CQA sites. They are divided into four groups: personal features, community features, textual features, and contextual features.

- *Personal features* These features are based on users' characteristics. Personal features include the activity of an answer's owner, including the number of answers given by the user, the number of questions asked by the user, the rank achieved by the user in the community, and the user's grade level.

- *Community features* These features are based on the response of the community to a user's answers, such as how many thanks they received or how many bans they received. Furthermore, we also consider the social connectivity of users in the community. In Brainly, users can make friends and exchange information. The friendships can be placed on a graph where users are nodes and the edge between two nodes represents the friendship. We extract several features about their connection—such as the number of friends—the clustering coefficient of a user and their egonet (aka, the friends of friends). The clustering coefficient (CC_i) of a user measures how closely their neighbors form a clique, defined as

$$CC_i = \frac{\# \text{ of triangles connected } i}{\# \text{ of connected triples centered on } i}. \quad (1)$$

Higher values mean that this user and their friends form a stronger connection. We denote $d_i = |N(i)|$ as the number of friends of users i , where $|N(i)|$ denotes the set of neighbors of i . Average degree of neighborhood is defined as

$$\bar{d}_N(i) = \frac{1}{d_i} \times \sum_{j \in N_i} d_j. \quad (2)$$

We also used the egonet information of users. These features can represent the structure of users connection. Furthermore, these features are all computed locally, which is scalable and efficient. Computing the community features is an *almost* linear time algorithm, taking *almost* $O(n \log n)$, where n is the number of nodes in graph.

- *Textual features* These features are based on answer content, such as length and format. We also check whether users use LaTeX for typing, since many answers provided in topical areas concerning mathematics and physics are easier to read if LaTeX is used. Furthermore, we measure the text readability based on two popular indexes: automated readability index (ARI) and Flesch reading ease score of answer (FRES) [17]. The ARI measures what grade level should understand the text, which is measured by

$$4.71 * \frac{\# \text{ of characters}}{\# \text{ of words}} + 0.5 * \frac{\# \text{ of words}}{\# \text{ of sentences}} - 21.43. \quad (3)$$

The FRES index measures the readability of the document. Higher FRES scores indicate the text is easier to understand. FRES index is calculated as

$$206.8 - 1.01 * \frac{\# \text{ of words}}{\# \text{ of sentences}} - 84.6 * \frac{\# \text{ of syllables}}{\# \text{ of words}} \quad (4)$$

Table 1 Lists of features on educational CQA are classified into four groups of features: personal, community, textual, and contextual*Personal features*

Number of answers given (n_answers)

Number of questions asked (n_questions)

Ranking of users (rank_id)

Grade level of users (u_grade)

Community features

Number of thanks that user received (thanks_count)

Number of warnings that user received (warns)

Number of spam reports that user received (spam_count)

Number of friends in community (friends_count)

Clustering coefficient in friendship network (cc)

Average degree of neighborhood (deg_adj)

Average CC of friends (cc_adj)

Size of ego network of friendship (ego)

Number of outgoing edges in ego network (ego_out)

Number of neighbors in ego network (ego_adj)

Textual features

Length of the answer (length)

Readability of the answer (ari)

Flesch reading ease score of the answer (fres)

Format of the answer (well_format)

Using advance math typing: LaTeX (contain_tex)

Contextual features

Grade level of question (q_grade)

Grade difference between answerer & question (diff_grade)

Rank difference between answerer & asker (diff_rank)

Similarity between answer and question (sim)

Device used to type the answer (client_type)

Duration to answer (time_to_answer)

Typing speed (typing_speed)

The abbreviations of features are in brackets

- *Contextual features* These features include the question's grade level, the device types used to answer the question, the similarity between answer and question, how long it took to type the answer, and the typing speed. The typing speed measures how many words the user types per second. The devices let us know whether the participant used a computer or a mobile device to answer. In order to compute the similarity between the answer and the question, we treat the answer and question as two vectors of words. The cosine similarity between these two vectors returns the similarity between them. Value 0 means that there are no common words between them. We believe that no common words between the answer and the question might indicate unrelated answers.

Table 2 Lists of features on focused CQA (Stack Overflow) are classified into four groups of features: personal, community, textual, and contextual*Personal features*

Number of answers given (n_answers)

Number of questions asked (n_questions)

Reputation of users (answerer_rep)

Community features

Number of up-votes that user received (up_votes)

Number of down-votes that user received (down_votes)

Number of friends in community (friends_count)

Clustering Coefficient in friendship network (cc)

Average degree of neighborhood (deg_adj)

Average CC of friends (cc_adj)

Size of ego network of friendship (ego)

Number of outgoing edges in ego network (ego_out)

Number of neighbors in ego network (ego_adj)

Textual features

Length of the answer (length)

Readability of the answer (ari)

Flesch reading ease score of the answer (fres)

Format of the answer (well_format)

Using advance math typing: LaTeX (contain_tex)

Contextual features

Rank difference between answerer & asker (diff_rep_users)

Similarity between answer and question (sim)

Duration to answer (time_to_answer)

The abbreviations of features are in brackets

Similarly, we have a list of features that pertain to subject-focused CQA. Though they are slightly different, they are essentially equivalent to those that apply to general CQA. For example, the ranking of users in Brainly is similar to the reputation of users in Stack Overflow. In focused CQA such as Stack Overflow, the good content is up-voted, and the user can achieve a respected reputation by providing high-quality content. In many CQA, friendship connections are not available. Thus, we can create the virtual friendship network [13]. In particular, there is a connection between users if they interact via asking–answering activity. Furthermore, some features, such as the device type or typing speed, are not available in the public data set. Fortunately, the results suggest that these features are not very important and have low prediction power.

Building training set In order to build the training data, we extracted features for each answer as given in Table 1. These can also be divided into two types of features. (i) *Immediate features* are the length, device type, typing speed, and similarity between answers and questions. These features are extracted immediately when the answer is posted. (ii) *History features*, such as the number of thanks and the number of

answers given, can be built beforehand and updated whenever these features change. Thus, when a new answer is posted, we can immediately extract all the proposed features, allowing our method to work in real time. Next, we describe three classifiers used in our study.

3.2 Classification

Because our framework could operate with almost any classification model, we compared the performance of different models in this study. In particular, we tested the classification algorithms below [4]. Let $X = x_1, x_2, \dots, x_n$ be the list of features. The list of classification algorithms is summarized as:

- *Logistic regression (log-reg)* Log-reg is a generalized linear model with sigmoid function

$$P\left(Y = 1|X = \frac{1}{1 + \exp(-b)}\right), \quad (5)$$

where $b = w_0 + \sum(w_i \cdot x_i)$, w_i are the inferred parameters from regression.

- *Decision trees* The tree-based method is a nonlinear model that partitions features into smaller sets and fits a simple model into each subset. The decision tree includes two-stage processes: tree growing and tree pruning. These steps stop when a certain depth is reached or each partition has a fixed number of nodes.
- *Random forest (RF)* RF is an average model approach [5,16], and we use a bag of 100 decision trees. Given a sample set, the RF method randomly samples data and builds a decision tree. This step also selects a random subset of features for each tree. The final outcome is based on the average of these decisions. The pseudo-code of RF is described in Algorithm 1. There are some advantages of RF. When building each tree in Step 4, RF randomly selects a list of features and a subset of data. Thus, RF can avoid the over-fitting problem of the decision tree. Furthermore, each tree can be built separately, which makes distributively computing the trees extremely easy.

Figure 3 summarizes the architecture of our method. In the framework, textual features and contextual features can be quickly calculated as soon as a new answer is posted. Personal and community features are extracted from the history database. After querying personal and contextual features, some features related to a user's activities (e.g., number of answers increased over time, etc.) are also updated accordingly.

Next, we will describe the data sets used in our study, as well as some characteristics of users in online learning communities.

Algorithm 1 Pseudo-code of random forest algorithm

Input:

- A set of training input $T = \{(X_i, y_i)\}, i = 1, \dots, n$.
- Number of trees N_{trees}
- A new feature vector X_{new}

Output: the prediction outcome of X_{new}

```

1: for  $i = 1 : N_{\text{trees}}$  do
2:   Randomly select a subset of training  $T_{\text{rand}} \subset T$ 
3:   Build the tree  $h_i$  based on  $T_{\text{rand}}$ 
4:   In each internal node of  $h_i$ , randomly select a set of features and
     split the trees based on these selected features
5: end for
6:  $\text{Pred}(X_{\text{new}}) = \sum_{i=1}^{N_{\text{trees}}} h_i(X_{\text{new}})$ 
7: return  $\text{Pred}(X_{\text{new}})$ 

```

4 Data sets and characterization of the data

In this section, we describe the data sets used in our study and several statistics characteristics of these data sets. In our study, we used data from two sites: Brainly and Stack Overflow.

4.1 Brainly: an educational CQA

Overview Brainly.com is an online Q&A for students and educators with millions of active users. In our study, we use the data from two markets: the USA and Poland (PL). Table 3 describes some characteristics of these data sets. Our study uses both deleted and approved answers. Brainly requires high-quality answers, meaning that moderators delete incorrect answers, incomplete answers, or spam posts. A moderator is an experienced user who has contributed significantly to the community. Established in 2013, the US market—currently consisting of 60,000 active users—is emerging in Brainly. In contrast, Brainly has been used since 2009 in PL and thus boasts an established market of 760,000 active users in that country. The posts in Brainly are divided into three levels (grades): primary, secondary, and high school. There is no detail category for each level.

Ranking of users Brainly uses a gamification-based feature that illustrates how actively users participate in answering questions. In the current Brainly system, there are seven hierarchical ranks, from beginner to genius. Users can advance through these categories in two ways: receiving a high number of points on their answers, and having their answers selected as “best” by askers. This ranking mechanism is similar to other CQA sites such as Yahoo! Answers and Stack Overflow, which encourage users to contribute to the site in order to earn a high reputation.

Deleting answers in Brainly Brainly tries to maintain high-quality answers, and moderators are recruited to participate heavily in deleting questions. Only experienced users, such as moderators, are allowed to delete answers. Answers may be

Fig. 3 An overview of a framework proposed in the study. In the Build Model step, we collected different groups of features for existing answers and the quality of these answers based on the community feedback. Based on these labeled data, we can build the classifier. Then, we can apply the model in Apply Model step. In this step, when a new answer appears, we do not know the quality, but we can quickly compute the groups of features (personal, community, textual, and contextual features) for this new answer. The computed features are used as the input to the classifier, and the classifier will predict the quality of answers automatically

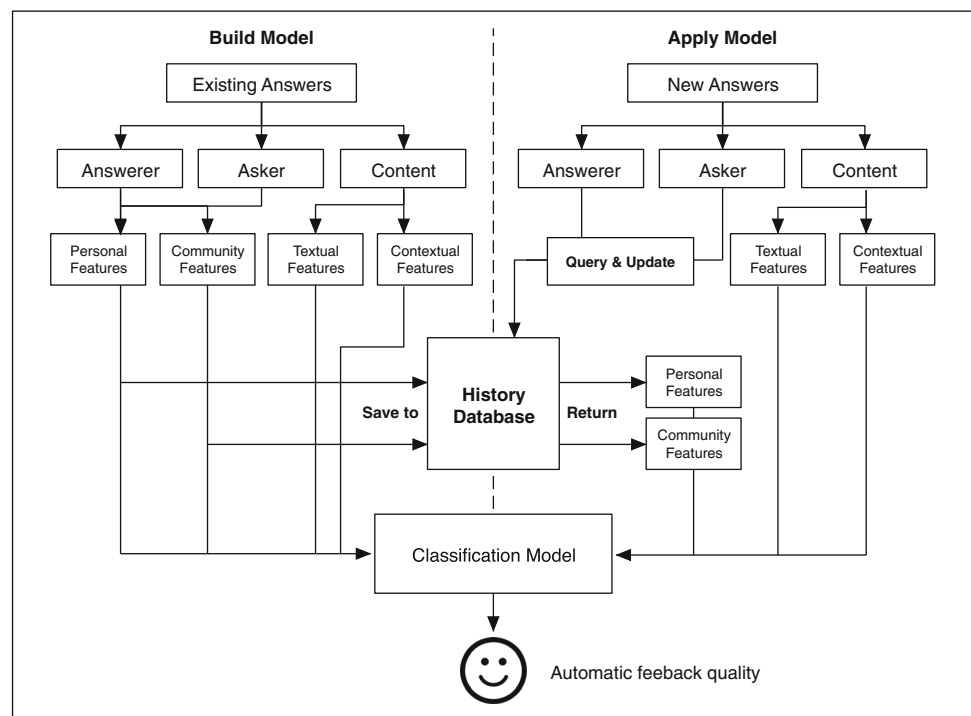


Table 3 Description of data

Site	Period	# of users	# of posts	# of answers
US	Nov'13–Dec'15	800 K	1.5 M	700 K
PL	Mar'09–Dec'15	2.9 M	19.9 M	10 M
Stack Overflow	Jul'08–Sep'14	3.4 M	21.2 M	7.7 M

US is the Brainly data in the USA market, and PL is the Brainly data in Poland

deleted if they are incomplete, incorrect, irrelevant, or spam. A significant portion of answers are deleted (30%) to maintain the site's high quality. But deleting this many answers is time-consuming and labor intensive. Furthermore, manual deletion might not be prompt, meaning that unsuitable content remains on the site until moderators have a chance to review it. Thus, developing an automatic mechanism to assess the quality of answers is a critical task.

Some popular reasons for deleting answers are listed in Table 4. There are many reasons answers are considered incorrect. In the majority of cases, posts are deleted due to answerers' insufficient knowledge.

Friendship in Brainly Users in this social CQA can make friends and exchange ideas and solutions. After joining the community, users can request to make friends with other users if their topics of interest are related. The friendship feature in Brainly is a new mechanism that encourages students to exchange ideas and solutions. In traditional CQA such as Yahoo! Answers and Stack Overflow, there is no formalized friendship connection. Figure 4 depicts the distribution of number of friends per user. We see that it follows the power law with long tail. Some users have many connections in

the community, while others make only a few contacts. We expect that users with many connections are more active and more committed to answering questions.

Activity in Brainly This is a free community. Anyone can contribute by asking questions, giving answers, giving thanks, and making friends. Due to the nature of the community, the contribution of each user is different and based on their interests and availability. Figure 5 shows the plot of the distribution of number of answers given per user. Again, this follows the power law with some very active users. Answering questions is a popular way for users to earn higher scores and increase their ranking in the community. Active users provide many answers to demonstrate that they are willing to devote their time to helping others. Answerers can also gain knowledge and trust from the community by answering a high volume of questions. Thus, these users may provide high-quality answers.

Subjects of interest The questions in Brainly are divided into different subjects/topics, such as Mathematics, Physics, etc. We examine how students participate in these topics between two countries. Figure 6 shows that students in both countries participate more in the topical areas of Mathematics, History,

Table 4 Popular reasons for deleting answers

Reason	Example
Lack of explanation (or uninformative)	Question: Out of 32 students in a class, 5 said they ride their bikes to school. Based on these results, how many of 800 students in the school ride their bikes to school? Approved answer: $(5/32) = (x/800)$. Set up the equation, multiply both sides by 800, $(4000/32) = x$, simplify $x = 125$ Deleted answer: 125 students
Mistakes	Question: Simplify $3(7x + 2y)$? Approved answer: It is $21x + 6y$. You would just multiply $7 * 3 = 21$ then u would bring down the x, then bring down the addition sign then multiply $3 * 2 = 6$ Deleted answer: $3(9x)$ is the simplified
Incomplete	Question: The two sociologists who referred to society as being a kind of living organism were (A) Auguste Comte; Emile Durkheim; (B) Karl Marx; Max Weber; (C) Auguste Comte; Herbert Spencer; (D) Emile Durkheim; Max Weber Approved answer: Auguste Comte; Emile Durkheim Deleted answer: AUGUSTE COMTE
Too vague	Question: How do u add and subtract fractions with different denominator? Approved answer: You first have to get both of the denominators to be the same. For example: if you're adding $\frac{1}{2} + \frac{1}{4}$, you need to multiply the $\frac{1}{2}$ by $\frac{2}{2}$ to get the fraction to be $\frac{2}{4}$ so you can add it with $\frac{1}{4}$ Deleted answer: You need to make them have a common denominator
Plagiarism	Question: Why shouldn't we bath after having lunch or dinner? Approved answer: Due to the cold water flowing on the body, the blood circulation increases around the body near the skin. This reduces the blood flow in the stomach... Deleted answer: Following ingestion food the splanchnic circulation or blood circulation to the viscera or the different parts and organs concerned with food...
Grammatical linguistic errors	Question: What color is given for desert in map? Approved answer: Brown Deleted answer: The answer is (BROWN)

Answers are deleted for diverse reasons. High-quality answers are approved on Brainly

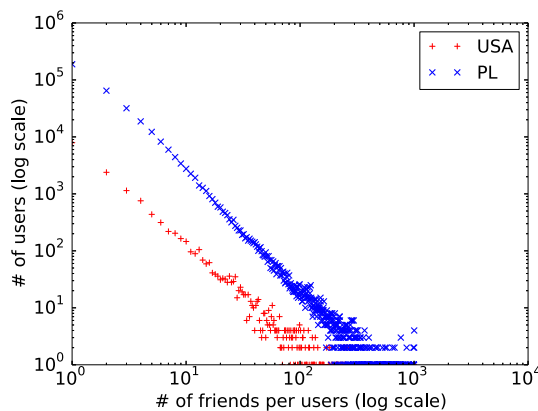


Fig. 4 Distribution of number of friends per user in log–log scale. The number of friends follows power law. Some users make a lot of friends in this community. US is the Brainly data in the US market, and PL is the Brainly data in Poland

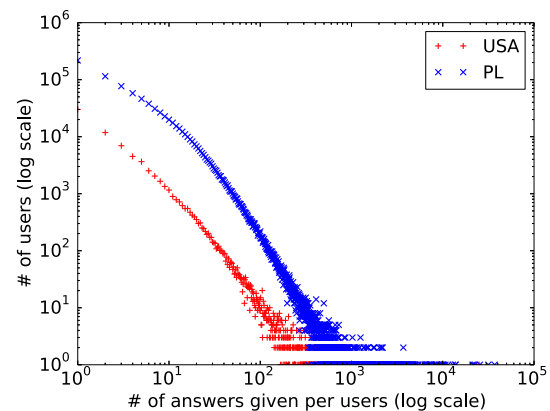


Fig. 5 Distribution of number of answers given per user. A small fraction of users answer a lot of questions, while many users answer a few number of questions

and English. The percentage of posts on Mathematics in the USA is significantly higher than in Poland (42% vs. 35%). This might indicate that students in the USA need more help with Mathematics.

The readability of answers We want to see whether approved answers are more readable—or more clear—than deleted answers. We use ARI to measure answers' readability. Findings indicate that the ARI of approved answers is 6.9 ± 3.1 ,

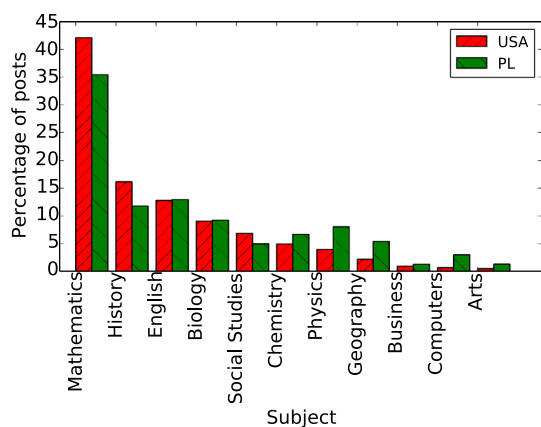


Fig. 6 Percentage of posts in different subjects. Both countries are similar, and students are most active in discussing Mathematics, History, and English (US is the Brainly data in the US market, and PL is the Brainly data in Poland)

while the ARI of deleted answers is 5.1 ± 3.2 . Similarly, the FRES indexes of deleted and approved answers are 69.9 ± 23.1 and 62.2 ± 22.5 , respectively. A higher FRES value means that an answer is easier to read. We see that the standard deviation is large for both indexes due to the diversity of content. We used t test when we wanted to check if a given feature for high-quality answers and low-quality answers is significantly different from each other. We conducted a t test and saw that the difference is significant with $p = 0.05$. This difference exists because many answers in primary and secondary levels are deleted. In general, the answers in primary and secondary levels are easy to read.

Quality of the answers of experienced and new users We examine the quality of answers from both new users and experienced users. We examine the deletion rate of answers based on the ranking of users. Figure 7 shows the plot of the rate of answers deleted from differently ranked users. We see that low-ranked users have a very high rate of deletion. Because Brainly supports education, the community expects correct answers. Even incomplete answers are deleted. We see that many intermediate users' (such as rank 3 or rank 4 users) answers are deleted. This demonstrates that Brainly maintains a very high standard to ensure quality answers.

We also use a popular CQA in our study. In the next section, we will describe Stack Overflow, the most popular CQA in programmer community.

4.2 Stack Overflow: a focused CQA

Overview Stackoverflow.com is another CQA specifically focused on the field of programming. All the questions and answers posted in Stack Overflow must be related to programming thus. Users in Stack Overflow can engage in a

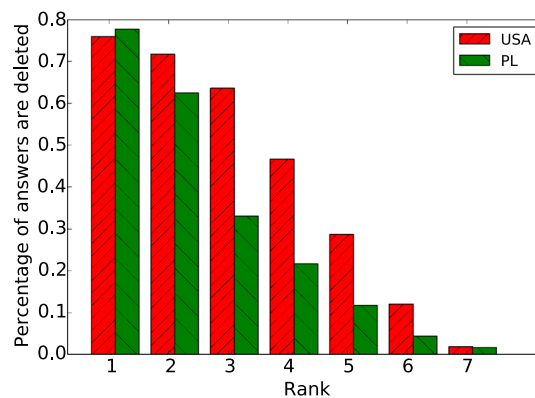


Fig. 7 Percentage of answers deleted versus rank level. Rank 1 is beginner while rank 7 is genius user. Highly ranked users have fewer deleted answers due to their experience. High deletion rate shows the site's answer requirements are very strict (US is the Brainly data in the US market, and PL is the Brainly data in Poland)

Table 5 Score system in Stack Overflow

Events	Change in reputation
Answer is up-voted	+ 10
Question is up-voted	+ 5
Answer is accepted	+ 15
Answer is down-voted	- 2
Question is down-voted	- 2
Answer win bounty	+ Bounty amount
Offer bounty	- Bounty amount
Answer marked as spam	- 100

wide range of activities and earn reputations by providing high-quality questions and answers.

Table 3 lists the characteristics of the data set used in our experiment. The data include information on all users and all posts from Stack Overflow since its creation in 2008 until September 2014. There are more than 21 million posts in this data set. Each question in Stack Overflow has its own tags, which are considered the post topics. Each question has one to five tags.

Users in Stack Overflow can engage in different activities such as posting questions, giving answers, voting for the best answer, and up-voting or down-voting a post. They can also earn a reputation by posting high-quality questions and answers. The table shows how this reputation is calculated. Table 5 summarizes the score system in Stack Overflow. In order to maintain high-quality posts and encourage high-quality posting, the score system is very rewarding and but strict.

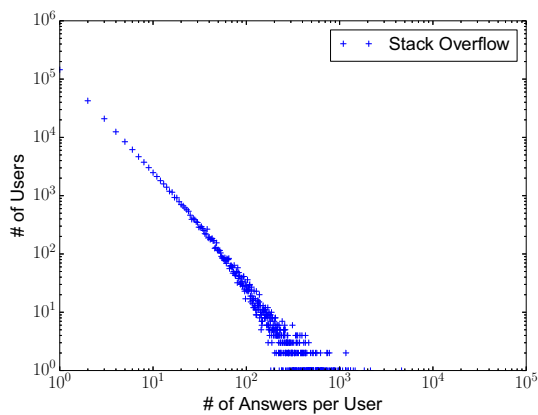


Fig. 8 Number of answers given per user in Stack Overflow. The contribution follows the power law

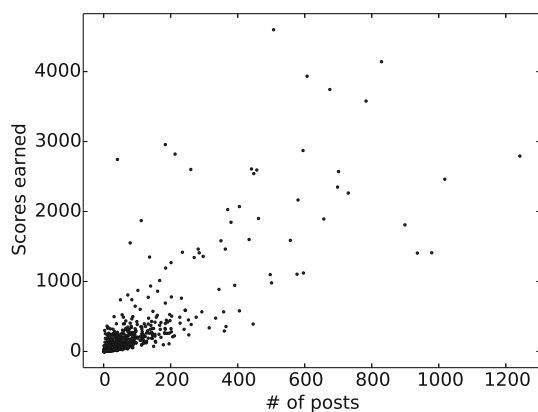


Fig. 9 Number of posts versus the score earned by 1000 randomly selected users who stay at least 1 year in Stack Overflow. A more active user can earn a higher score

4.2.1 Number of posts

Figure 8 shows the plot of the number of answers per user. It shows that a few users are very active and contribute a majority of the content in Stack Overflow. This reflects the general observation that many users in CQA only answer a few questions, while a small number of users are very active.

4.2.2 Number of posts versus user reputation

Figure 9 shows the relationship between 1000 randomly selected users' number of posts and earned scores. The figure demonstrates that if a user contributes more frequently, they can earn a higher score. This relationship, however, is not linear because the quality of the post significantly affects a user's score. For example, if a user posts 1000 mid- or low-quality posts, they may earn the same score as another user who contributes less than 100 high-quality posts. But in general, a user can only become a top contributor if they are active in the community.

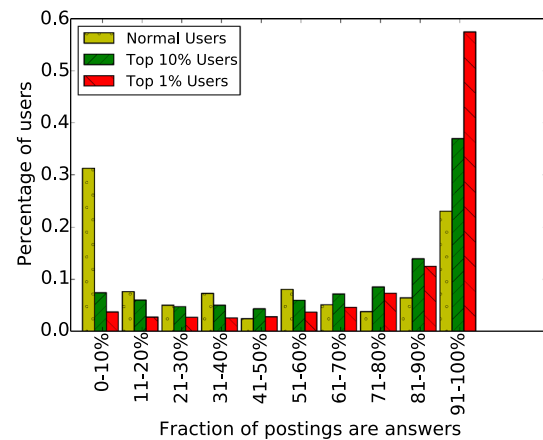


Fig. 10 Comparing users' question–answering behavior. The x-axis is the percentage of posts that are answers. Most top contributors' posts are answers. (US is the Brainly data in the US market, and PL is the Brainly data in Poland)

4.2.3 Question–answering behavior versus user reputation

Users in CQA can contribute to the community and earn their reputation by posting useful questions and answers. We want to see whether there are differences between normal users and top contributors. Figure 10 shows the distribution for different types of users. We see that top contributors' posts are often answers. For example, in the group of top contributors, 60% of these users list more than 90% of their posts as answers. In contrast, more than 30% of normal users have less than 10% of their posts as answers. This indicates that many normal users may only join a CQA community to find answers.

In the previous section, we describe and analyze some characteristics of data set used in our study. In the next section, we will describe the experiment setup, the main results, and the discussion of the results.

5 Experiments and results

In this section, we will describe our experimental setup, highlight the main results, and provide a discussion around these experiments and findings.

5.1 Experimental setup

We will describe the settings. The common settings refer to the setting to all data sets. The **Common settings**: To compare the performance of classifications, we used different classification algorithms with different sets of features. In the default setting, we used the random forest of 100 decision trees. In the evaluation, we randomly selected 200,000 answers in each data set to validate our framework's accu-

racy. We used tenfold cross-validation to select parameter classification with 70–30% training and testing set. In order to compare the efficacy, we examined the accuracy, F1-score, confusion matrix, and area under curve.

Settings based on a particular data set

- *Brainly's data sets* Our framework predicts whether an answer is deleted or approved by the community moderators.
- *Stack Overflow data set* We use the community feedback as a metric to measure the quality of the answer. The high-quality answer gets the positive score, and the low-quality answer receives the negative score. The score is based on the number of up-votes and down-votes.

5.2 Main results

5.2.1 Accuracy

Accuracy is defined as the percentage of correctly classified answers. Figure 11 shows the plot of the accuracy of applying random forest (RF) to different groups of features. *PF*, *CmF*, *TF*, and *CtF* denote the results when our frameworks used personal features, community features, textual features, and contextual features, respectively. *All* presents the accuracy when using every feature. The results demonstrate that personal features and community features are more useful in predicting answer quality. This makes sense because good users normally provide good answers. The textual features have less prediction value due to the complexity of the site's content. Furthermore, our classifier achieves a very high accuracy of more than 78.5% in all data sets. Given the complexity of answers within the community, these results are very encouraging.

5.2.2 F1-score

We also measure *F1* score, which considers both precision and recall. Precision is the fraction of instances that are relevant, while recall is the fraction of relevant instances that are retrieved. The value of *F1* is defined as

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

Figure 12 shows that using all features achieves the highest *F1* score, which is more than 84% in both Brainly data sets (USA and PL) and 78.5% in Stack Overflow. High *F1* scores demonstrate that our method can achieve both high precision and recall. Similar to accuracy, the results suggest that personal features and community features are the most important in the model.

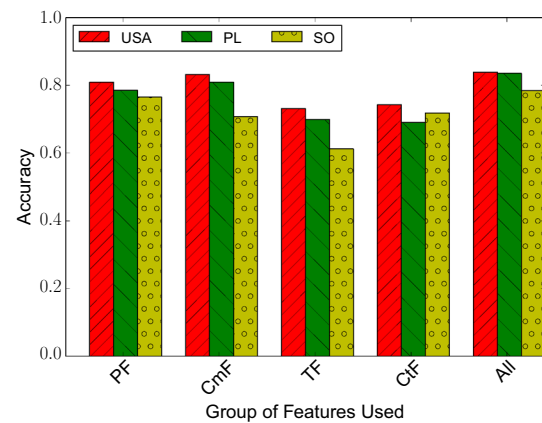


Fig. 11 Accuracy of using different groups of features. *PF*, *CmF*, *TF*, and *CtF* denote the results when our frameworks used personal features, community features, textual features, and contextual features, respectively. *All* means using all features. *PF*, *CmF* are more useful in predicting the quality of answers (random forest is the classifier used; US is the Brainly data in the USA market, PL is the Brainly data in Poland, and SO is the Stack Overflow data)

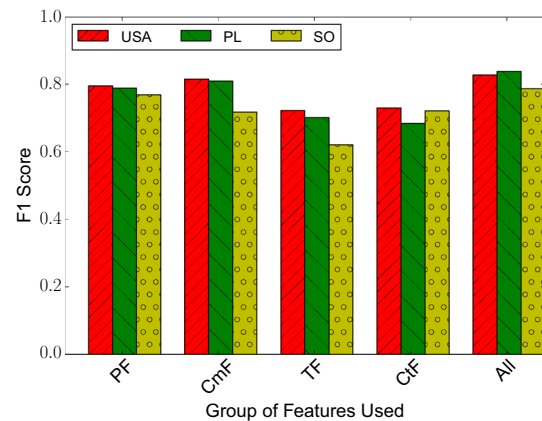


Fig. 12 Comparison of the *F1* score (higher is better) when using different groups of features. Random forest is the classifier used. High *F1* score shows that our method achieves high value in both precision and recall. Again, personal features and community features are more important in the model (US is the Brainly data in the US market, PL is the Brainly data in Poland, and SO is the Stack Overflow data)

The performances in the Brainly data sets are higher than those in Stack Overflow (83.5% vs. 78.5%). Several reasons explain this discrepancy. First, Stack Overflow provides a public data set that does not contain every feature. Furthermore, the content in Stack Overflow is more complicated than that in Brainly because Stack Overflow is used widely by professionals, such as programmers, with specialized knowledge.

5.2.3 Comparing different classifiers

Table 6 compares the accuracy when applying different classification algorithms. We see that random forest outperforms

Table 6 Comparison of the accuracy of different classifiers

Classification	USA (%)	PL (%)	Stack Overflow (%)
Logistic regression	79.1	76.8	73.1
Decision trees	78.2	77.1	72.9
Random forest	83.9	83.5	78.5

Random forest (bag of 100 trees) outperforms logistic regression and decision trees

logistic regression and decision trees. This is due to the nonlinear relationship between features and answer quality. Furthermore, random forest also randomly selects different sets of features to build trees, which avoids over-fitting in classification. Random forest is also an efficient algorithm that can work well on large data sets. Our experiment was conducted on a machine with 2.2 GHz quad-core, 16 GB of RAM. It was implemented in Python code on a data set consisting of 200,000 answers. The experiment took 34 seconds to train the model and less than 1 millisecond to predict each answer. Training is one-time cost. It implies that our framework can determine the quality of answers in real time. Thus, our suggestion is to use random forest as a classifier in a real system.

5.2.4 Feature importance

In this section, we measure which features are more important. In order to determine this, we use a permutation test to remove the features and measure the accuracy of out-of-bag (OOB) samples. The important features will substantially degenerate the accuracy. Figure 13 reports the importance of different features used in our study. The three most important features are the number of thanks users receive, the amount of spam reported, and the similarity between answers and questions. We believe that some features strongly correlate with

quality but ultimately appeared less important, such as device type or using LaTeX when typing. For example, participants using mobile devices to submit their answers may make more mistakes, or a participant using LaTeX markup might indicate their high experience with certain topics. Unfortunately, there were only a few answers that were posted from mobile devices or typed in LaTeX (less than 10%). Thus, these features lost their prediction value. In the Stack Overflow data, the social connection feature is less important because the site facilitates virtual connections. As we mentioned before, there is no formal friendship in Stack Overflow. In a smaller community such as a forum, the virtual connection is stronger and contains more information [13]. But because CQA is a large community, creating a virtual connection based on asking–answering interactions has a less prediction value.

We also consolidated the features into feature categories for easier interpretation. Figure 13 shows the plot the importance of different group of features. Again, personal features and community features are more important in the model.

5.2.5 Features selection

One possible concern is whether features selection can improve the performance of our method. The general idea of features selection is to remove features that have no correlation with the outcome, or to remove two similar features. In both cases, such features cause over-fitting in the prediction. In random forest, we already randomly select features when building the trees. In particular, Step 4 in Algorithm 1 selects random features to build the trees. Furthermore, the number of features in our study is not large. Thus, features selection is unnecessary and does not help improve accuracy.

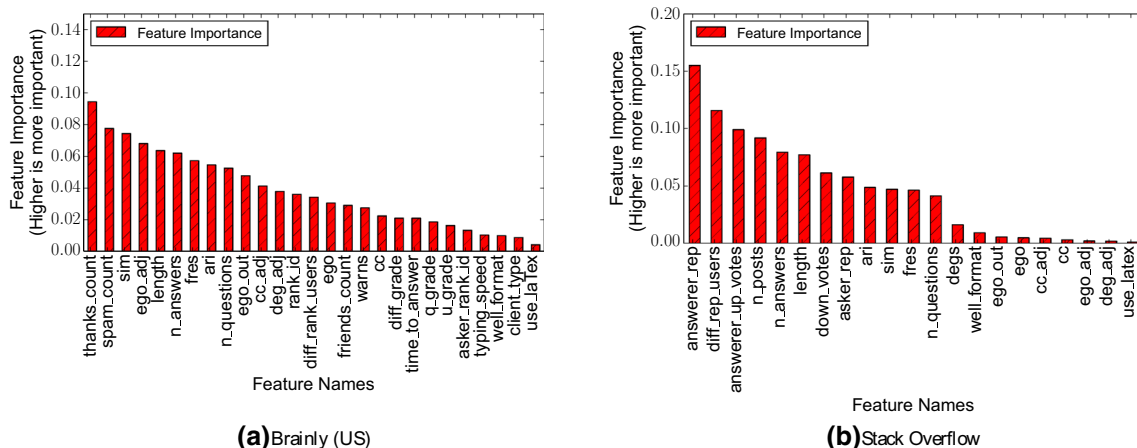


Fig. 13 Measure of important features (higher is more important). Most important features are *number of thanks—up-votes*, *reputation*, *number of spams*, *similarity between question and answer*, and so on. Tables 1 and 2 list the notations of used features

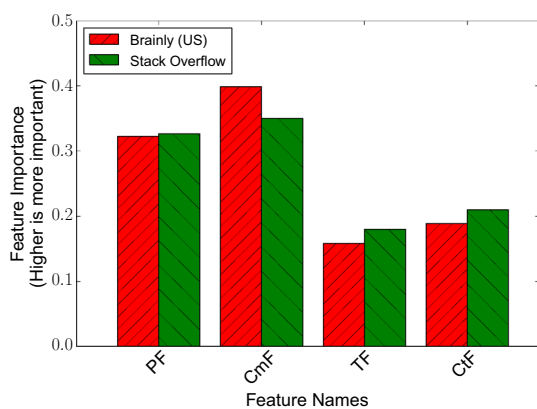


Fig. 14 Measure of important features by aggregating into different groups (higher is more important). Again, personal features and community features are more important in the model

Table 7 Confusion matrix for predicting answer quality

Actual value	Prediction outcome		Total (%)
	Deleted (%)	Approved (%)	
<i>a. Brainly—USA</i>			
Deleted	90.1	9.9	100
Approved	22.4	77.6	100
<i>b. Brainly—Poland</i>			
Deleted	81.5	18.5	100
Approved	14.5	85.5	100
<i>c. Stack Overflow</i>			
Negative	77.7	22.3	100
Positive	20.6	79.4	100

5.2.6 High-quality answers and low-quality answers

We discuss which is more difficult to detect: high-quality or low-quality answers. Table 7 examines the confusion matrix that describes how answers are wrongly classified in different data sets. We see that detecting deleted questions achieves higher accuracy than detecting approved answers in the US. This is so because many answers in the US market are answered by newcomers, who do not often satisfy the high-quality criteria established by this CQA community. In the PL market, there is no difference due to a well-established community and the fact that the majority of the participants are experienced users. Similarly, there is no difference in Stack Overflow data since Stack Overflow is also a well-established community.

5.2.7 Receiver operating characteristic (ROC)

We also evaluate the ROC of the approved answers for both data sets. The ROC denotes the classification's ability to find the correct high-quality answers with different thresholds.

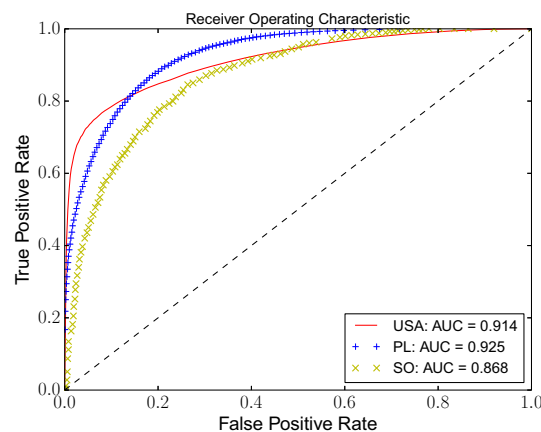


Fig. 15 Area under ROC curve for our frameworks is above 0.9 in both Brainly data sets and is 0.86 in Stack Overflow data set

The curve in Fig. 15 plots the true positive rate against the false positive rate. We see that the area under ROC is higher than 0.91 in both data sets. In the real deployment, we can set different thresholds to select the approved answers based on various requirements. For example, the administrators of the site might believe that 17% is insufficient and require that the automatic assessments not make mistakes with a rate of more than 0.05. Figure 15 shows that if the false positive rate is 0.05, the true positive rates of the US and PL are 0.73 and 0.62, respectively. Otherwise, we can detect a majority of the approved answers with a small error rate. The rest of the answers are considered borderline entities, which are hard to differentiate between good and bad. Under these circumstances, we can still take advantage of moderators and askers to provide another evaluation of the questions or answers.

It is more difficult to detect the quality of answers in Stack Overflow since this is a focused CQA where the questions and answers are much more difficult and complex than those in educational CQA. In Stack Overflow, the content is more varied in nature. The covered topics (e.g., Python programming, Hadoop) are also constantly evolving and there are new topics emerging. Thus, it is more difficult to develop a framework that can assess the quality of Stack Overflow answers automatically. For example, Stack Overflow has thousands of topics. In our framework, the reputation of users is an important feature. An experienced user in Java does not guarantee high skill in machine learning. In contrast, Brainly has far fewer categories/topics that are bound in the type of content (e.g., Physics, Math, Biology), giving a better concentration of the content that a user generates. In all cases, our model significantly reduces humans' workload.

6 Discussion

Asking a question for learning is not a new phenomenon within the area of information seeking. It is an innate and purposive human behavior to satisfy an information need via searching techniques [8], and information and knowledge received through an asker’s questioning behavior may become meaningful if the information acquired helps solve their problematic situations [42]. In recent years, new information and communication technologies have emerged to develop novel ways for users to interact with information systems and experts in order to seek information. These new resources include digital libraries and virtual reference tools, as well as CQA services that allow users to be both consumers and producers of information.

According to Ross et al. [28], librarians and experts in both face-to-face and virtual reference environments engage in a process of negotiating an asker’s question. This helps identify an asker’s information need and allows them to construct a better question that will receive higher-quality answers. However, this process of question negotiation may not occur in the context of CQA, which significantly impedes an asker’s ability to receive satisfactory answers. Identifying what constitutes the content quality of information generated in CQA (or for that matter, any online repository with user-generated content) can be critical to the applicability and sustainability of such digital resources.

When it comes to CQA in educational contexts, seeking and sharing high-quality answers to a question may be more critical since question–answering interactions for educational answers are likely to solicit factual and verifiable information, in contrast to general-purpose CQA services where advice and opinion-seeking questions are predominant [7]. Thus, evaluating and assessing the quality of educational answers in CQA is important not only for improving user satisfaction, but also for supporting students’ learning processes. In more complex CQA that pertains to focused and difficult topics, communities heavily rely on moderators, who often find themselves with inordinately high work loads. Thus, the accuracy of our framework is encouraging, as it suggests that an automated process can alleviate stress and improve content within CQA communities.

In this work, we attempted to investigate an educational and a focused CQA, by examining a series of textual and non-textual answer features in order to identify levels of content quality among the answers. The study first attempted to identify a list of content characteristics that would constitute answer quality. In the second step, we applied these features in order to automatically assess the quality of answer. The results showed that personal features and community features are more robust in determining answer quality. Most of these features are available and feasible to compute in other CQA

sites, making our approach applicable to the wider community.

Furthermore, the efficacy and efficiency of our method makes it possible to implement within the real system. In our experiment on a standard PC, it takes less than one millisecond to return the prediction. It also only takes less than one minute to train the model with 200,000 answers from Brainly. However, the training step is a one-time cost and can be accomplished using distributed processing. By applying this technique to the real system, we believe that we can reduce the number of deleted answers by giving a warning immediately before a user submits a response to the community. Furthermore, the approach can approve high-quality answers and subsequently significantly reduce an asker’s wait time.

Most of the previous work applied logistic regression in order to evaluate the quality of answers. Our work showed that a “wisdom of the crowd” approach, such as random forest, can significantly improve the accuracy of assessing answer quality due to a nonlinear relationship between the features and the quality of answers. The results, for instance, showed that longer answers are more likely to be approved compared to deleted answers. However, it is also worth noting that very lengthy answers might signal low-quality answers, such as confusing or spam answers. Even though the current study focused on evaluating the quality of answers in terms of educational information on CQA for online learning in particular, the study also suggests an alternative way of how general CQAs’ answer quality could be investigated by our method—i.e., a “wisdom of the crowd” approach—in order to improve the accuracy of quality answer assessment. Moreover, in terms of practical implications of users’ interactions for content moderation on CQA, the findings may propose a variety of features or tools (e.g., detecting spams, trolling, plagiarism, etc.) that support content moderators in order to develop a healthy online community in which users may be able to seek and share high-quality information and knowledge via question–answering interactions.

There are also limitations to our work. For example, it can only detect high- and low-quality answers. It would be helpful if we could provide suggestions to improve the overall quality of these answers. We believe this is challenging but highly rewarding work that might require a significant effort to examine answer meaning. Furthermore, our approach was heavily based on the communities’ past interactions and thus has limited applicability to a newly formed CQA community.

7 Conclusion

The work described here focused on answer quality in both educational and subject-focused CQAs. Educational CQAs support young students by seeking and sharing educational

information in different topics related to school subjects (e.g., Mathematics, Physics, etc.), while focused CQAs support building learning communities for professional users (e.g., computer languages, etc.). During the current work, we focused on improving the efficiency of community management in terms of facilitating users' ability to seek and share high-quality answers to a question. Traditional methods involving human assessments may not be sufficient due to the large amount of content available, as well as subjective assessments of answer quality; therefore, we propose a framework to automatically assess the quality of answers for these communities. In general, our framework integrated four different aspects of answers, such as personal features, community features, textual features, and contextual features. We presented here the first large-scale study on CQA for education and focused programming topics. Our method achieves high performance in all important metrics such as accuracy, F1 score, and area under ROC curve. Furthermore, the experiment demonstrates that our method is highly efficient and can work well in a real-time system.

We conducted a large-scale study of two popular CQAs: the educational Brainly, which spanned two major markets, and the programming-focused Stack Overflow. Our framework performs well in all data sets. In this study, we have access to all questions generated in Brainly, while the Stack Overflow data contained that which was publicly released. The content in Stack Overflow is more complicated than that in Brainly because Stack Overflow is used widely by professionals, such as programmers, with specialized knowledge. Because Stack Overflow contains more complex content, our model demonstrated slightly lower accuracy when applied to its data set. However, we find that personal features and community features are more robust in assessing the quality of answers in an online community. The textual features and contextual features are less robust due to the diversity of users and content in these communities. For example, a user with a high reputation has the higher chance to give a good answer than a new user with a lower reputation. For both communities (i.e., Brainly, Stack Overflow), we found that the higher the user's reputation, the better the quality of answers given. The length and the effort spent on the answer are also important features. Since the user spends more time to think and type the answer, his or her answer may be easier to understand and may be accepted by the community. In contrast, knowing the users using mobile device or computer does not help much. Using LaTeX will make the answers related to mathematics or physics more readable. Unfortunately, only a small number of answers used LaTeX. Thus, the feature that checks whether a user uses LaTeX in answer has a minimal capability in predicting the quality of answers. Furthermore, all features used in this study can be computed easily, which makes the framework's implementation feasible.

In our study, we, however, did not provide details of what determines the quality of answers. Although it was found that some answers are deleted due to their low quality, the current study did not include any specific reasons of deletion; however, potential reasons may include insufficient information provided in answers, any user behaviors against community policies (e.g., using bad words, bullying other users, etc.). Furthermore, the current approach does not help the user (asker/answerer) with the *rightness* or *wrongness* of the answer. Checking for the factual accuracy of an answer is beyond the scope of this work. We understand that correctness of an answer is an important element of its quality, and when a method is developed to automatically detect it, the proposed framework here should be able to integrate that method for even better quality assessment. Thus, future works may focus on expanding the current framework to identify what constitutes low quality of answers in educational and focused CQAs, to support informing how users would be able to develop and share better answers to satisfy an asker's need in online environments. In addition, as CQA is a user-driven community where users voluntarily create all contents, future works should also include how to assist users to create good content, which further increases overall CQA quality. Understanding latent features related to users who struggle to contribute to creating good content would help improve users' experiences on seeking and sharing information in CQA communities as well as make online learning more effective.

Acknowledgements The work reported in this paper is supported by the US Institute of Museum and Library Services (IMLS) Grant #LG-81-16-0025. A portion of the work reported here was possible due to funds and data access provided by Brainly. We are also grateful to Michal Labedz and Mateusz Burdzel from Brainly for their help and insights into the topics discussed in this work. We are also thankful to Stack Overflow for sharing their data.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: WWW, pp. 665–674 (2008)
2. Aritajati, C., Narayanan, N.H.: Facilitating students' collaboration and learning in a question and answer system. In: CSCW Companion, pp. 101–106 (2013)
3. Berlingerio, M., Koutra, D., Eliassi-Rad, T., Faloutsos, C.: Network similarity via multiple social theories. In: ASONAM, pp. 1439–1440 (2013)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, Berlin (2006)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
6. Choi, E., Borkowski, M., Zakoian, J., Sagan, K., Scholla, K., Ponti, C., Labedz, M., Bielski, M.: Utilizing content moderators to investigate critical factors for assessing the quality of answers on brainly, social learning Q&A platform for students: a pilot study. In: ASIST, pp. 69:1–69:4 (2015)

7. Choi, E., Kitzie, V., Shah, C.: Developing a typology of online Q&A models and recommending the right model for each question type. In: ASIST, pp. 1–4 (2012)
8. Choi, E., Shah, C.: User motivation for asking a question in online Q&A services. *J. Assoc. Inf. Sci. Technol.* **67**(5), 1182–1197 (2016)
9. Cole, R.A.: *Issues in Web-Based Pedagogy: A Critical Primer*. Greenwood Press, Westport (2000)
10. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Exploiting user feedback to learn to rank answers in Q&A forums: a case study with stack overflow. In: SIGIR, pp. 543–552 (2013)
11. Dalip, D.H., Lima, H., Gonçalves, M.A., Cristo, M., Calado, P.: Quality assessment of collaborative content with minimal information. In: JCDL, pp. 201–210 (2014)
12. Dror, G., Maarek, Y., Szpektor, I.: Will my question be answered? Predicting “question answerability” in community question-answering sites. *ECML/PKDD* **8190**, 499–514 (2013)
13. Ganu, G., Marian, A.: Personalizing forum search using multidimensional random walks. In: ICWSM, pp. 140–149 (2014)
14. Gazan, R.: Social Q&A. *JASIST* **63**, 2301–2312 (2011)
15. Gollapalli, D., Mitra, P., Giles, C.L.: Ranking experts using author-document-topic graphs. In: JCDL, pp. 87–96 (2013)
16. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, Berlin (2009)
17. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Air Station Memphis (1975)
18. Le, L.T., Eliassi-Rad, T., Tong, H.: MET: a fast algorithm for minimizing propagation in large graphs with small eigen-gaps. In: SDM, pp. 694–702 (2015)
19. Le, L.T., Shah, C.: Retrieving rising stars in focused community question-answering. In: ACIIDS, pp. 25–36 (2016)
20. Le, L.T., Shah, C., Choi, E.: Evaluating the quality of educational answers in community question-answering. In: JCDL, pp. 25–36 (2016)
21. Le, L.T., Shah, C., Choi, E.: Bad users or bad content? Breaking the vicious cycle by finding struggling students in community question-answering. In: CHIIR, pp. 165–174 (2017)
22. Levy, A.Y., Rajaraman, A., Ordille, J.J.: Querying heterogeneous information sources using source descriptions. In: VLDB, pp. 251–262 (1996)
23. Liu, Y., Bian, J., Agichtein, E.: Predicting information seeker satisfaction in community question answering. In: SIGIR, pp. 483–490 (2008)
24. Momeni, E., Tao, K., Haslhofer, B., Houben, G.-J.: Identification of useful user comments in social media: a case study on flickr commons. In: JCDL, pp. 1–10 (2013)
25. Noer, M.: One man, one computer, 10 million students: how khan academy is reinventing education. In: *Forbes* (2013)
26. Pelleg, D., Yom-Tov, E., Maarek, Y.: Can you believe an anonymous contributor? On truthfulness in yahoo! answers. In: SOCIALCOM-PASSAT, pp. 411–420 (2012)
27. Preece, J., Nonnecke, B., Andrews, D.: The top five reasons for lurking: improving community experiences for everyone. *Comput. Hum. Behav.* **20**(2), 201–223 (2004)
28. Ross, C., Nilsen, K., Dewdney, P.: *Conducting the Reference Interview: A How-to-do-it Manual for Librarians*. NealSchuman, New York (2002)
29. Shah, C., Kitzie, V.: Social Q&A and virtual reference—comparing apples and oranges with the help of experts and users. *JASIST* **63**, 2020–2036 (2012)
30. Shah, C., Oh, S., Oh, J.S.: Research agenda for social Q&A. *Libr. Inf. Sci. Res.* **31**(4), 205–209 (2009)
31. Shah, C., Pomerantz, J.: Evaluating and predicting answer quality in community QA. In: SIGIR, pp. 411–418 (2010)
32. Shah, C., Radford, M., Connaway, L., Choi, E., Kitzie, V.: How much change do you get from 40\$? Analyzing and addressing failed questions on social Q&A. In: ASIST, pp. 1–10 (2012)
33. Srba, I., Bielikova, M.: Askalot: community question answering as a means for knowledge sharing in an educational organization. In: CSCW Companion, pp. 179–182 (2015)
34. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers on large online QA collections. In: ACL, pp. 719–727 (2008)
35. Surowiecki, J.: *The Wisdom of Crowds*. Anchor, New York City (2005)
36. Suryanto, M.A., Lim, E.P., Sun, A., Chiang, R.H.L.: Quality-aware collaborative question answering: methods and evaluation. In: WSDM, pp. 142–151 (2009)
37. Szymczak, V.: Github and Stackoverflow in Technical Recruitment. <http://sourcingrecruitment.info/2015/05/github-and-stackoverflow-in-technical-recruitment/>. Accessed 30 Dec 2016
38. Tan, C.H., Agichtein, E., Ipeirotis, P., Gabrilovich, E.: Trust, but verify: predicting contribution quality for knowledge base construction and curation. In: WSDM, pp. 553–562 (2014)
39. Tess, P.A.: The role of social media in higher education classes (real and virtual)—a literature review. *Comput. Hum. Behav.* **29**, A60–A68 (2013)
40. Wang, G., Gill, K., Mohanlal, M., Zheng, H., Zhao, B.Y.: Wisdom in the social crowd: an analysis of quora. In: WWW, pp. 1341–1352 (2013)
41. Yang, L., Bao, S., Lin, Q., Wu, X., Han, D., Su, Z., Yu, Y.: Analyzing and predicting not-answered questions in community-based question answering services. In: AAAI, pp. 1273–1278 (2011)
42. Yang, S.: Information seeking as problem-solving using a qualitative approach to uncover the novice learners’ information-seeking process in a perseus hypertext system. *Libr. Inf. Sci. Res.* **19**(1), 71–92 (1997)
43. Yao, Y., Tong, H., Xie, T., Akoglu, L., Xu, F., Lu, J.: Joint voting prediction for questions and answers in CQA. In: ASONAM, pp. 340–343 (2014)
44. Yao, Y., Tong, H., Xu, F., Lu, J.: Predicting long-term impact of CQA posts: a comprehensive viewpoint. In: SIGKDD, pp. 1496–1505 (2014)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

International Journal on Digital Libraries is a copyright of Springer, 2019. All Rights Reserved.