Evaluating the Quality of Educational Answers in Community Question-Answering

Long T. Le
Department of Computer Science
Rutgers University
longtle@cs.rutgers.edu

Chirag Shah
School of Communication and Information
Rutgers University
chirags@rutgers.edu

Erik Choi
Brainly
erik.choi@brainly.com

ABSTRACT

Community Question-Answering (CQA), where questions and answers are generated by peers, has become a popular method of information seeking in online environments. While the content repositories created through CQA sites have been used widely to support general purpose tasks, using them as online digital libraries that support educational needs is an emerging practice. Horizontal CQA services such as Yahoo! Answers and vertical CQA services such as Brainly are aiming to help students improve their learning process through answering their educational questions. In these services, receiving high quality answer(s) to a question is a critical factor not only for user satisfaction, but also for supporting learning. However, the questions are not necessarily answered by experts, and the askers may not have enough knowledge and skill to evaluate the quality of the answers they receive. This could be problematic when students build their own knowledge base by applying inaccurate information or knowledge acquired from online sources. Using moderators could alleviate this problem. However, how moderators evaluate the quality of answers may be inconsistent since it is based on their subjective assessments. Employing human assessors may also be insufficient due to a large amount of content available on a CQA site. To address these issues, we propose a framework for assessing the quality of answers automatically. This is achieved by integrating different groups of features - personal, community-based, textual, and contextual - to build a classification model and determine what constitutes answer quality. To test this evaluation framework, we collected more than 10 million educational answers posted by more than 3 million users on Brainly's United States and Poland sites. The experiments conducted on these datasets show that the model using random forest achieves more than 83% accuracy in identifying the high quality of answers. In addition, the findings indicate that personal features and community-based features have more prediction power in assessing answer quality. Our approach also achieves high values on other key metrics such as F1-score and Area under ROC curve. The work reported here can be useful for many other contexts where providing automatic quality assessment in a digital repository of textual information is paramount.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; E.1 [**Data Structures**]:

General Terms

Algorithms; Design; Performance; Experimentation

Keywords

Community Question-Answering (CQA); Answer Quality; Features

1. INTRODUCTION

The Internet and the World Wide Web (WWW) have become critical and ubiquitous information tools that have changed the way people share and seek information. 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 [15], users have adopted various types of such online information sources in order to seek and share information. These include Wikis, forums, blogs, and community question-answering (CQA). CQAs are one example of a new means of information seeking through users sharing information and knowledge in virtual environments.

According to Gazan [11], 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 asking and answering a question. Within CQA, there are other elements, such as commenting and voting, that encourage social interactions for seeking and sharing information. Because of the fast growth of CQA's popularity, a rich body of research has been conducted in order to understand the variety of content and user behaviors in question-answering interactions within the context of CQA. Shah et al. [22] describe that previous studies based on user content have focused on content type, quality, and formulation, while studies focusing on user behaviors attempted to understand the motivations for asking and answering a question on CQA.

Many of the initial CQA platforms, such as AnswerBag (the first one in the US), were developed to support general purpose information seeking. They are referred to as the horizontal CQA services. Then, other sites were deployed for more specific tasks - vertical CQA. One type of specific task or purpose is online learning. In education, students not only use the Internet to look for new material but can also exchange ideas and knowledge. The advent of CQA has assisted students greatly in sharing knowledge in virtual

environments. As CQA in education is an emerging field, educators hope that they may be able improve learning capability and experience with the help of communication and information technologies.

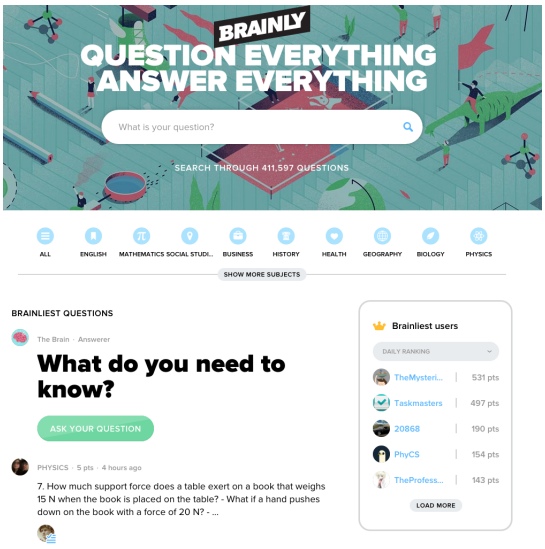

Figure 1: Brainly's homepage in United States.

To further this push for employing CQA services and content for educational purposes, we attempt to examine Brainly,[1] one of the largest CQA services specifically targeted at education. Brainly is a leader in online social learning networks for students and educators with millions of active users. It has approximately 60 million monthly unique visitors as of January 2016 and is available in 35 countries, including the United States, Poland, Russia, Turkey, Brazil, France, and Indonesia. Figure 1 shows the homepage of Brainly in the United States.

CQA is a user driven community where all contents including questions and answers are generated by community members. Thus, content quality is an important aspect in retaining existing users and attracting new members. The quality of information for educational purposes is even more important. For example, students who use the CQA to ask questions about homework problems could be mislead by wrong answers. This is an especially problematic issue for struggling students. Thus, quality assessment is a critical aspect. At the moment, traditional CQAs depend on human judgment to evaluate content quality. There are several drawbacks of this mechanism, including subjective (and possibly biased) assessments employed by the assessors, seeming difficulty in recruiting such evaluators, and the time it could take for human assessors to go through ever-increasing content in CQA sites. The work reported here addresses these concerns by providing a new framework for assessing content quality. Our specific contributions are as follows.

- Empirical study: this is the first large scale study to investigate the quality of answers of in an emerging CQA for education.

- Propose a framework to assess the answers automatically. Our frameworks extract 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 83% in both data sets. Our method also achieves high values on other key metrics such as F1 score and Area under ROC curve.

- Examine 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 our paper is organized as follow: Section 2 discusses the background and a few of the related works. The framework is described in Section 3. Section 4 presents the data sets used in our study. We present the results and discussion of our method in Section 5. Finally, the conclusion and future work are presented in Section 7.

2. BACKGROUND AND RELATED WORK

2.1 Community Question-Answering (CQA)

Community Question-Answering (CQA) services have become popular places for Internet users to look for information. Some popular CQA such as Yahoo! Answers or Stack Overflow attract millions of users. CQA takes advantage of *Wisdom of the Crowd*, the idea that everyone knows something [26]. Users can contribute to the community by asking questions, giving answers, and voting for the posts. Most activities are moderated by humans.

Several works have investigated user interest and motivation for participating in CQA [19], [29]. Adamic et al. [1] studied the impact of CQA. In this work, the authors analyzed questions and clustered them based on the question's 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. The works also examined the best answers by using some basic features such as the length and pasts answers. Shah et al. [21] 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 more accurately highlight potential strategies for matching question askers with question answerers.

2.2 CQA for Online Learning

In recent years, online learning has collapsed time and space [8], 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, social media (e.g., Facebook, Twitter, etc.) has attracted attention for empirical investigations conducted in order to understand the effectiveness of higher education [28]. 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 [18]. Additionally, even though most CQAs are mainly focused on either general topics (e.g., Yahoo! Answers, WikiAnswers, etc.) and/or professional topics (e.g, Stack Overflow, etc.) to seek and share information, new CQAs have emerged to help students participate in question-answering interactions to share educational information for online learning. Examples of such educational CQA include Chegg [2] and Brainly. Brainly specializes in online learning for students (i.e,

[1]http://brainly.com

[2]https://www.chegg.com

middle school, high school) through asking and answering activities in 16 main school subjects (e.g., English, mathematics, biology, physics, etc.) [5].

2.3 Quality Assessment in CQA

Since most contents in CQA are generated by users who actively seek and share information with other users, the content quality is a critical factor to the success of the community. Therefore, 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 for their educational information needs. High quality content is the best way to retain existing users and attract new members [16]. However, assessing the quality of posts in CQA is a difficult task due to the diversity of contents and users.

Examining the quality of answers can be divided into three types of problems: *(i)* finding the best answer, *(ii)* ranking the answers, and *(iii)* measuring the quality of answers. For example, Shah el al. [23] looked for the best answers in Yahoo! Answers by using 13 different criteria. Ranking answers is a useful task when a question receives multiples answers. These works focus more on the similarity between answer and question [25]. Surynato et al. [27] took advantage of the expertise of an asker and an answerer to rank the answers. In this work, the authors also recognized that different users are experts in different subjects and used this understanding to rank the answers. Recent work also showed the potential of using graphs in ranking users [12], but it is not clear how to rank answers based on users. The last type of problem focuses on regression-related problems such as predicting how many answers a question will get or how much posts can attract the interest from the community. Researchers are interested in predicting whether certain questions in CQA will be answered and how many answers a question will receive [30, 10]. This research used features such as asker history, the length of question, and the question category to predict the answeribility of the question. Shah et al. [24] studied why some questions remain unanswered in CQA. Particularly, this work explored why fact-based questions often fail to attract an answer. Momeni et al [17] applied machine learning to judge the quality of comments in online communities, revealing that social context is a useful feature. Yao et al. [32] examined the long-term effect of the posts in Stack Overflow by developing a new scalable regression model. Dalip et al. [9] tried to reduce the number of features in collaborative content, however the number of reduction was not significant. Furthermore, applying feature selection can solve the issue with many features such as over-fitting.

Our work is close to measuring the quality of answers. This research uses past question-answering interactions and current question and/or answering activities in order to predict the quality of new answers automatically. The framework incorporates different groups of features including personal features, community features, textual features, and contextual features.

3. EXAMINING THE QUALITY OF AN ANSWER

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

Formal definition:

Given:

- a set of users $U = \{u_1, u_1, ..., u_n\}$
- a set of posts $P = Q \cup A$,

  Q is the set of questions $Q = \{q_1, q_2, ..., q_{m1}\}$, and
  A is the set of answers $A = \{a_1, a_2, ..., a_{m2}\}$
- a set of interactions $I = \{i_1, i_2, ..., i_{m3}\}$ (such as giving thanks, 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 the history and information of users in the community, 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 training models.

3.1 Feature Extraction

In order to classify the quality of answers, we build a list of features for each answer. Table 1 lists the features used in our study. The features are divided into four groups: Personal Features, Community Features, Textual Features, and Contextual Features.

- Personal Features: These features are based on the characteristics of users. Personal features include the activity of an answer's owner such as the number of of answers given by the user, the number of questions asked by the user, the rank that user achieved 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 he got or how many bans he got. Furthermore, we also consider the social connectivity of users in the community. In Brainly, users can make friends and exchange information. The friendships create 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, clustering the coefficient of user and their ego-net (aka, the friends of friends). The clustering coefficient (CC_i) of a user measures how close their neighbors form a clique, defined as

$$CC_i = \frac{\#\ of\ triangles\ connected\ i}{\#\ of\ connected\ triples\ centeredon\ i} \quad (1)$$

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

$$d_N(i) = \frac{1}{d_i}\sum_{j \in N_i} d_j \quad (2)$$

  These features incorporate four social theories, which are Social Capital, Structural Hole, Balance, and Social Exchange [2]. Furthermore, these features are all computed locally which is salable and efficient.*is salable the word you want to use here?* This is an *almost* linear time algorithm, taking *almost* $O(n \log n)$, where n is number of node in graph.
- Textual Features: These features are based on answer content such as the length of answers and the format of answers. We also check whether users use Latex for typing since many answers provided in mathematics and physics topic areas are

easier to read if Latex is used. Furthermore, we also measure the readability of the text based on two popular indexes: automated readability index (ARI), and Flesch reading ease score of answer (FRES) [14]. The ARI measures what grade level should understand the text, which is measured by

$$\alpha * \frac{\#ofcharacters}{\#ofwords} + \beta * \frac{\#ofwords}{\#ofsentences} - \gamma \quad (3)$$

where α, β, γ are 4.71, 0.5, 21.43 respectively based on empirical study. The FRES index measures the readability of the document, and is calculated as

$$\alpha' - \beta' * \frac{\#ofwords}{\#ofsentences} - \gamma' * \frac{\#ofsyllables}{\#ofwords} \quad (4)$$

Again, α', β', γ' are calculated based on empirical study and are equal to 206.8, 1.01, 84.6 respectively. Higher FRES scores indicate the text is easier to understand.

- Contextual Features: These features contain some contextual features such as the question's grade level, the device types used to answer the question, the similarity between answer and question, duration to 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.

Building training set: In order to build the training data, we extracted features for each answer as seen 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 number of answers given can be built beforehand and be updated whenever this feature changes. Thus, when a new answer is posted, we can extract all proposed features immediately, which means our method can work in real time. Further details about these settings are described in Section 5. Next, we describe three classifiers used in our study.

3.2 Classification

Since our framework could use almost any classification model, we compared the performance of different models in this study. In particular, we tested the below classification algorithms [3]. Let $X = x_1, x_2, ..., x_n$ is the list of features. The list of classification algorithms are summarized as:

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

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

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

Personal Features
Number of answers given
Number of question asked
Ranking of users
Grade level of users
Community Features
Number of thanks received
Number of warnings that user received
Number of spam reports that user received
The ranked achieved
Number of friends in community
Clustering Coefficient in friendship network
Average degree of neighborhood
Average CC of friends
Size of ego-network of friendship
Number of outgoing edges in ego-network
Number of neighbors in ego-network
Textual features
The length of answer
The readability of answer (ARI)
The Flesch Reading Ease Score of answer (FRES))
The format of answer
Using advance math typing (latex)
Contextual features
The grade level of question
The grade difference between answerer & question
The similarity between answer and question
Device used to type answer
Duration to answer
Typing speed

Table 1: List of features are classified into four groups of features: Personal, Community, Textual, and Contextual.

- 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 [13, 4] 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 computing the trees distributively extremely easy.

Figure 2 summarizes the architecture of our method. In the framework, textual features and contextual features can be calculated quickly at the moment when 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 and 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, ..., n.
- Number of trees N_{trees}
- A new feature vector X_{new}

Output: the prediction outcome of X_{new}

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

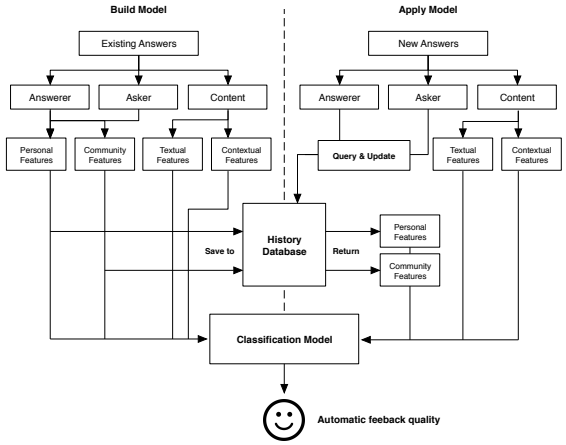

Figure 2: An overview of a framework proposed in the study.

4. DATASETS AND CHARACTERIZATION OF THE DATA

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 United States (US) and Poland (PL). Table 2 describes some characteristics of these data sets. In our study, we use two types of answers: deleted answers and approved answers. Brainly requires high quality answers. Thus incorrect answers, incomplete answers, or spam posts are deleted by moderators. A moderator is an experienced user who has contributed significantly to the community. The United States is an emerging market for Brainly, which was established in 2013. In contrast, Poland is a well-established market where Brainly has been used since 2009. 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-related feature that illustrates how actively users participate in answering questions. In the current Brainly system, there are seven hierarchical ranks, from Beginner to Genius, that users can advance to based on how many points they receive through answering a question, as well as how many of their answers are selected as the best answer by an asker. This 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.

Site	Period	# of Users	# of Posts	# of Answers
US	Nov '13 to Dec '15	800 K	1.5 M	700 K
PL	Mar '09 to Dec '15	2.9 M	19.9 M	10 M

Table 2: Description about data.

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

Friendship in Brainly: Users in this social CQA can create friendships and exchange ideas and solutions. The friendship feature in Brainly is a new mechanism, which encourages students to exchange ideas and solutions. In traditional CQA such as Yahoo! Answers and Stack Overflow, there is no formalized friendship connection. Figure 3 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 connections. We expect that users with many connections are more active and more committed to answering questions.

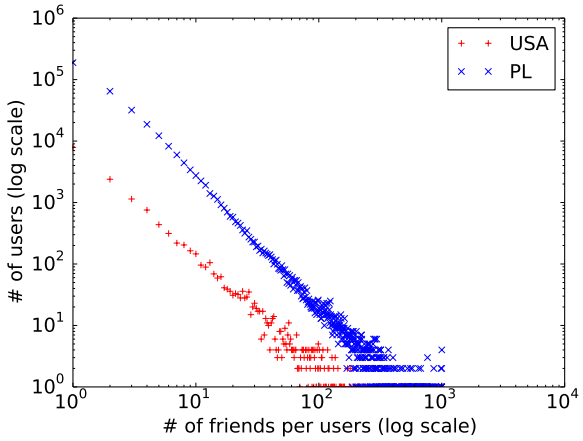

Figure 3: Distribution of number of friends per user in log-log scale. The number of friends follows power law. Some users make a lot of friends in this community.

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 4 plots 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. Giving many answers shows that these active users are willing to devote their time to helping others. Answering a high number of questions also helps answerers gain knowledge and trust from the community. Thus, answers from these users could have high quality.

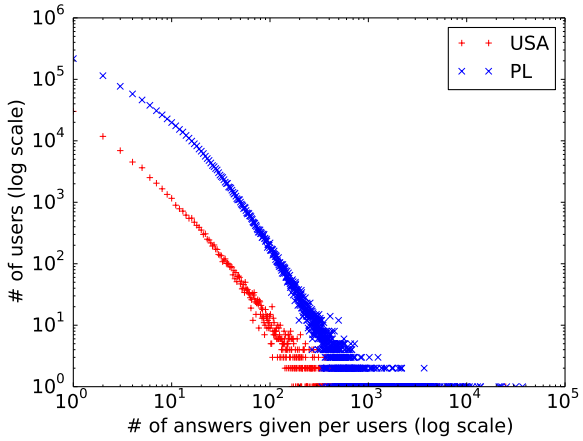

Figure 4: Distribution of number of answers given per users. A small fraction of users answer a lot of questions while many users answer a few number of questions.

Subject of interest: The questions in Brainly are divided into different topics such as mathematics, physics, etc. We examine how students participate in these topics between two countries. Figure 5 shows that student in both countries participate more in the topic areas of mathematics, history, and English. The percentage of posts on mathematics in the United States is significantly higher than in Poland (42% vs. 35%) This might indicate that students in the US need more help with mathematics.

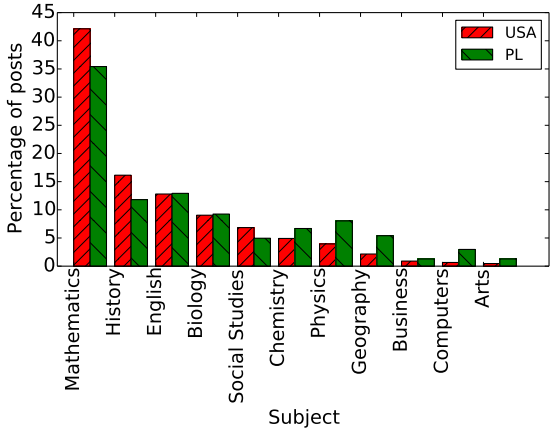

Figure 5: Percentage of posts in different subjects. Both countries are similar and students are most active in discussing mathematics, history, and english.

The readability of answers: We want to see whether the approved answers are more clear to read than deleted answers. We use ARI to measure the readability of answers. It shows that the ARI of approved answers is 6.9 ± 3.1, the ARI of deleted answers is 5.1 ± 3.2. Similarity, the FRES indexes of deleted and approved are 69.9 ± 23.1 and 62.2 ± 22.5 respectively. The higher FRES value means that it is easier to read. We see that the standard deviation is large for both indexes due to the diversity of content. We did a t-test and see that the difference is significant with $p = 0.05$. The reason is many answers in primary and secondary levels are deleted. In general the answers in primary and secondary levels are easy to read.

Quality of experienced and newbie users: We examine the quality of answers of new users and experienced users. We examine the deletion rate of answers based on the ranking of users. Figure 6 plots the rate of answers deleted for differently ranked users. We see that low-rank users have a very high rate of deletion. Since Brainly is a CQA which supports eduction, the site expects correct answers. Even incomplete answers are deleted. We see that many answers of intermediate users (such as rank 3 or 4 users) are deleted. It shows that Brainly maintains a very high standard to ensure quality answers.

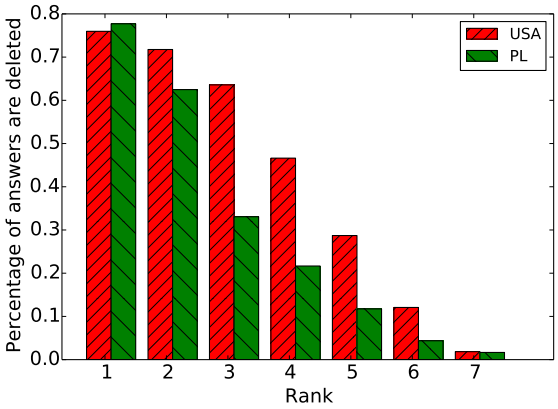

Figure 6: Percentage of answers deleted vs. rank level. Rank 1 is beginner while rank 7 is genius user. High ranked users have less deletion due to their experience. High deletion rate shows the site's answer requirements are very strict.

In the next section, we will describe the experiment set up, the main results, and the discussion about 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 compare the performance of classification using different classifications with different sets of features. In the default setting, we used the Random Forest of bag with 100 decision trees. In the evaluation, we randomly selected 200 K answers in each data set to validate the accuracy of our framework. We used 10-fold cross validation to select parameter classification with 70-30% training, testing set. In order to compare the efficacy, we examined the accuracy, F1-score, confusion matrix, and Area Under Curve.

5.2 Main results

5.2.1 Accuracy

Accuracy is defined as the percentage of answers classified correctly. Figure 7 plots the accuracy of using different groups of features when applying Random Forest. *PF, CmF, TF, CtF* denotes the results when our frameworks used personal features, community features, textual features, and contextual features respectively. *All* presents the accuracy when using all features in classification. The results show that personal features and community features are more useful in predicting the quality of the answer. The result makes sense since good users normally give good answers. The textual features have less prediction value due to the complexity of the site's content. We will examine the detail of each feature later. Furthermore, our classifier achieves very high accuracy - more than 83% in both markets. These results are very encouraging due to the complexity of answers in the community.

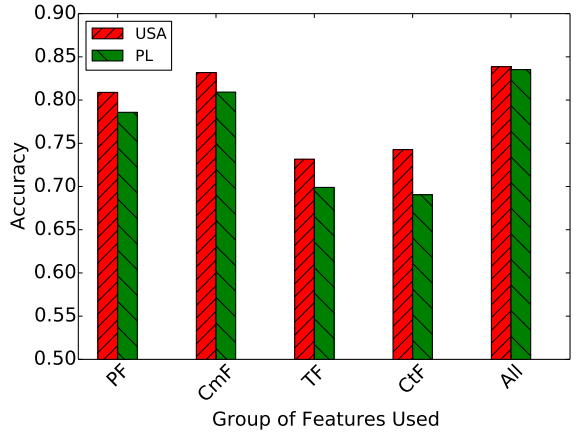

Figure 7: The accuracy of using different groups of features. *PF, CmF, TF, CtF* denotes the results when our frameworks used personal features, community features, textual features, and contextual features respectively. *All* means using all features. *PF, CmF* are more useful in predicting the quality of answers. (Random Forest is the classifier used.)

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

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

Figure 8 shows that using all features achieves the highest $F1$ score which is more than 84% in both data sets. High $F1$ scores show that our method can achieve both high precision and recall. The results are similar to accuracy, where personal features and community features are inferior features in the model.

5.2.3 Comparing different classifiers

Table 3 compares the accuracy when applying different classification algorithms. We see that Random Forest outperforms Logistic regression and decision trees. The reason is non-linear relation

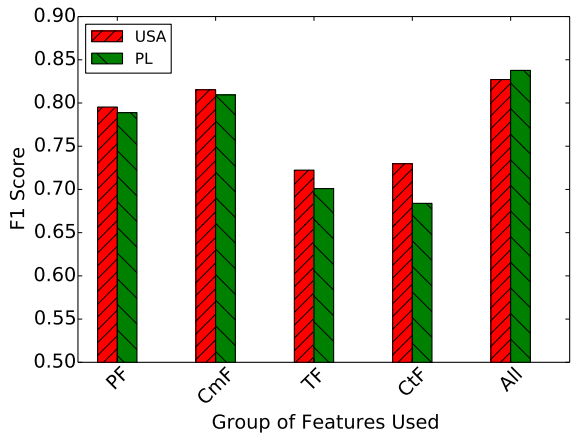

Figure 8: Compare the $F1$ score (higher is better) when using different groups of features. Random Forest is the classifier used. High F1 score shows that our method achieves high value in both precision and recall.

between features and the quality of answers. Furthermore, random forest also randomly selects different sets of features to build the trees, which avoids over-fitting in classification. Random Forest is also an efficient algorithm which can work well on large data sets. Our experiment was conducted on a machine with 2.2 GHz quad-core, 16 GB of RAM, implemented in Python code, on a data set is 200 thousand 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.

Classification	USA	PL
Logistic Regression	79.1%	76.8%
Decision Trees	78.2%	77.1%
Random Forest	**83.9%**	**83.5%**

Table 3: Compare the accuracy of different classifiers. Random forest (bag of 100 trees) outperforms logistic regression and decision tree.

5.3 Discussion

5.3.1 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-out-bag (OOB) samples. We suspect that the important features will degenerate the substantial enough accuracy. Figure 9 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. Some features are believed to have strong correlation with quality but are less important, such as device type or using Latex when typing. For example, participants using mobile devices to submit their answers may contain more mistakes, or a participant using Latex markup might indicate a user's high experience with certain topics. Unfortunately, there were only a few answers that were posted from mobile de-

		Prediction outcome		
		Deleted	Approved	Total
Actual value	Deleted	90.1%	9.9%	100%
	Approved	22.4%	77.6%	100%

a. United States

		Prediction outcome		
		Deleted	Approved	Total
Actual value	Deleted	81.5%	18.5%	100%
	Approved	14.5%	85.5%	100%

b. Poland

Table 4: Confusion matrix for predicting answer quality.

vices or typed in Latex (less than 10%). Thus, these features lost their prediction value.

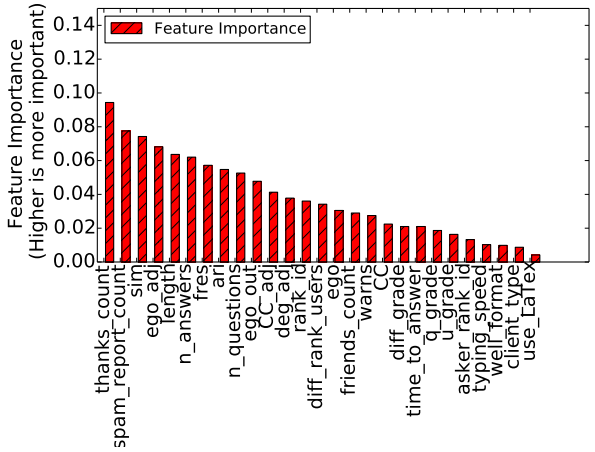

Figure 9: Measure features important (higher is more important).

5.3.2 *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 two similar features. In both cases, such features cause over-fitting in the prediction. In Random Forest, we already select features randomly 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.

5.3.3 *High quality answers and low quality answers*

We discuss which is more difficult to detect: high quality or low quality answers. Table 4 examines the confusion matrix which describes how answers are mis-classified in the US and PL. We see that detecting deleted questions achieves higher accuracy than detecting approved answers in the US. The reason is the many answers in the US market are answered by newcomers which does not 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.

5.3.4 *Receiver operating characteristic (ROC)*

We also evaluate the ROC of the approved answers for both data sets. The ROC denotes the ability of the classification to find the correct high quality answers with different thresholds. The curve in Figure 10 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 10 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 small error rate. The rest of the answers are considered borderline entities, which are hard to differentiate between good or bad. In this case, we can still take advantage of moderators and askers to evaluate the questions or answers again. In this case, the workload of humans is reduced significantly.

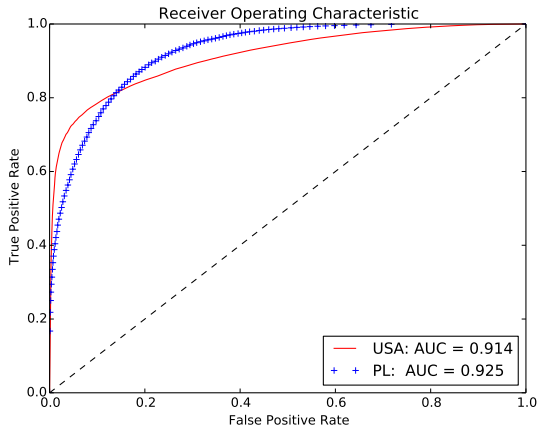

Figure 10: Area Under ROC curve for our frameworks are above 0.9 in both data sets.

6. DISCUSSION

Asking questions for the purpose of learning is not a new phenomenon within the area of information seeking. It is an innate and purposive human behavior to search information to satisfy a need [7], and information and knowledge received through an asker's questioning behavior may become meaningful in that the information acquired helps solve their problematic situations [31]. 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, virtual references, as well as CQA services where users are both consumers and producers of information.

According to Ross et al. [20], what librarians and experts in brick and mortar as well as virtual reference environments do is a process of negotiating an asker's question, which helps identify an asker's information need and allows him/her to construct a better question to receive high quality answers. However, this process of question negotiation does not occur in the context of CQA, which may cause significant issues for providing high quality answers. Identifying what constitutes the content quality of informa-

tion 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 for educational contexts, seeking and sharing high quality answers to a question may be more critical since question-answering interactions for educational answers is likely to solicit factual and verifiable information, contrasting with general-purpose CQA services where advice and opinion-seeking questions are predominant [6]. Thus, evaluating and assessing the quality of educational answers in CQA is important for not only improving user satisfaction, but also for supporting learning processes of the students.

In our current work, therefore, we attempted to utilize a series of textual and non-textual features of answers in order to identify a level of the content quality in a context of CQA for educational answers. The study first attempted to identify a list of content characteristics that would constitute the quality of answers. 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 the 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 make 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 the response to the community. Furthermore, the approach can approve high-quality answers, so that an asker's wait time can be reduced significantly.

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 improve the accuracy of assessing the quality of answers significantly. The reason is non-linearly relation between the features and the quality of answers. For example, the results showed that longer answers are more likely to be approved compared to deleted answers. But too long answers might be a signal of 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 would also suggest an alternative way of how the quality of answers in a context of general CQA would be investigated by the method, i.e., a "wisdom of the crowd" approach, in order to improve the accuracy of assessing the quality of answers. 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. Our work can only detect high and low-quality answers. It would be helpful if we could provide suggestions to improve the quality of these answers. We believe this is challenging work but a highly rewarding task, which might require a significant effort to examine answer meaning. Furthermore, our approach was based heavily on the community's past interactions, which has limited applicability to a newly-formed community.

7. CONCLUSION

In the current study, we focused on the quality of educational answers in CQA. Our work was motivated by a need to improve the efficiency of managing the community in terms of seeking and sharing high quality answers to a question. Since employing human assessments may not be sufficient due to the large amount of content available, as well as subjective assessments of answer quality, we propose a framework to assess the quality of answers automatically for these communities. In general, our framework integrates different aspects of answers such as personal features, community features, textual features, and contextual features. This is the first large scale study on CQA for education. Our method achieves high performance in all important metrics such as accuracy, F1 score, and Area under ROC curve. Furthermore, the experiment shows the efficiency of our method, which can work well in a real time system.

We find that personal features and community features are more robust in assessing the quality of answer in an online education community. The textual features and contextual features are less robust due to the diversity of users and content in these communities. Furthermore, all features used in this study can be computed easily which makes the framework's implementation feasible.

In future work, we plan to study struggling users in the community. We see that many answers were deleted due to low quality. But it is not clear the reason why the posts were deemed low quality. It might be due to a lack of knowledge on the part of the answerer, an arrogant attitude, or anti-social behavior. We believe that understanding the latent features can help struggling users, improve users' experiences, and make online learning more efficient.

8. ACKNOWLEDGEMENT

A portion of the work reported here was possible thanks to funds and data access provided by Brainly. We are also grateful to Mihal Labedz, Mateusz Burdzel, and Kent Scholla from Brainly for their help and insights into the topics discussed in this work.

9. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *WWW*, pages 665–674, 2008.

[2] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. In *ASONAM*, pages 1439–1440, 2013.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

[4] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[5] E. Choi, M. Borkowski, J. Zakoian, K. Sagan, K. Scholla, C. Ponti, M. Labedz, and M. Bielski. 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*, 2015.

[6] E. Choi, V. Kitzie, and C. Shah. Developing a typology of online q&a models and recommending the right model for each question type. *ASIST*, 49(1):1–4, 2012.

[7] E. Choi and C. Shah. User motivation for asking a question in online Q&A services. *JASIST*, In press.

[8] R. A. Cole. *Issues in Web-based pedagogy: A critical primer*. Greenwood Press, 2000.

[9] D. H. Dalip, H. Lima, M. A. Gonçalves, M. Cristo, and P. Calado. Quality assessment of collaborative content with minimal information. JCDL, pages 201–210, 2014.

[10] G. Dror, Y. Maarek, and I. Szpektor. Will my question be answered? predicting "question answerability" in community question-answering sites. In *ECML/PKDD*, volume 8190, pages 499–514, 2013.

[11] R. Gazan. Social q&a. *JASIST*, 63:2301–2312, 2011.

[12] S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking experts using author-document-topic graphs. JCDL, pages 87–96, 2013.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. 2009.

[14] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, 1975.

[15] A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *VLDB*, pages 251–262, 1996.

[16] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. SIGIR, pages 483–490, 2008.

[17] E. Momeni, K. Tao, B. Haslhofer, and G.-J. Houben. Identification of useful user comments in social media: A case study on flickr commons. JCDL, pages 1–10, 2013.

[18] M. Noer. *One Man, One Computer, 10 Million Students: How Khan Academy Is Reinventing Education*. Forbes, 2013.

[19] J. Preece, B. Nonnecke, and D. Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2):201 – 223, 2004.

[20] C. Ross, K. Nilsen, and P. Dewdney. *Conducting the reference interview: A how-to-do-it manual for librarians.* New York: NealSchuman, 2002.

[21] C. Shah and V. Kitzie. Social q&a and virtual reference - comparing apples and oranges with the help of experts and users. 63:2020–2036, 2012.

[22] C. Shah, S. Oh, and J. S. Oh. Research agenda for social Q&A. *Library & Information Science Research*, 31(4):205–209, 2009.

[23] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR*, pages 411–418, 2010.

[24] C. Shah, M. Radford, L. Connaway, E. Choi, and V. Kitzie. How much change do you get from 40$? analyzing and addressing failed questions on social q&a. In *ASIST*, pages 1–10, 2012.

[25] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers on large online qa collections. In *ACL-08*, pages 719–727, 2008.

[26] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[27] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang. Quality-aware collaborative question answering: Methods and evaluation. WSDM, pages 142–151, 2009.

[28] P. A. Tess. The role of social media in higher education classes (real and virtual) - a literature review. *Computers in Human Behavior*, 29:A60–A68, 2013.

[29] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: An analysis of quora. In *WWW*, pages 1341–1352, 2013.

[30] L. Yang, S. Bao, Q. Lin, X. Wu, D. Han, Z. Su, and Y. Yu. Analyzing and predicting not-answered questions in community-based question answering services. In *AAAI*, pages 1273–1278, 2011.

[31] S. Yang. Information seeking as problem-solving using a qualitative approach to uncover the novice learners' information-seeking process in a perseus hypertext system. *Library and Information Science Research*, 19(1):71–92, 1997.

[32] Y. Yao, H. Tong, F. Xu, and J. Lu. Predicting long-term impact of cqa posts: A comprehensive viewpoint. In *SIGKDD*, pages 1496–1505, 2014.