

Bad Users or Bad Content? Breaking the Vicious Cycle by Finding Struggling Students 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) services have become popular methods to seek and share information. In CQA, users with an information need, or askers, post a question that community members can answer. This question-answering process allows both askers and answerers to learn through the exchange of information. CQA services have also been widely used in the education domain, as some of such services are designed specifically for students' information seeking. However, due to insufficient knowledge, lack of experience, and other reasons, students often struggle in producing quality or even appropriate content. This low quality production causes their content to be flagged or deleted, further discouraging them from participating in the CQA process and instigating a vicious cycle of bad users and bad content. In an effort to break this cycle, the work reported here focuses on identifying users whose postings demonstrate a high deletion rate with a presumption that the *bad* content is an indication of a struggling student rather than a malicious user. In this work, experiments are conducted on a large student-oriented online CQA community to understand struggling students' behaviors. A framework is proposed to find these users based solely on their activities. Finally, community feedback (i.e. human judgment) such as moderator evaluation or community votes for good content is used to detect these users in the early stages of their respective struggles. To evaluate this framework, we used data from Brainly, a large educational CQA service that is used in two different markets with more than 3.7 million users and 10.7 million answers. The results show that the human judgment feature identifies early-stage struggling users with high accuracy. Identifying these struggling users (students) could help educators to determine suitable ways to help their students instead of presuming them to be *bad* users and cutting them off from the community.

1. INTRODUCTION

With the emergence of the Web and more specifically Web 2.0, a vast number of users have used interactive online communities in order to seek and share information. Examples of these online

community platforms include social networks, wikis, blogs, and community question-answering (CQA) sites. These platforms provide an unprecedented opportunity for users to enrich their minds and share knowledge. One popular way to ask a question is through CQA, where users often desire more personalized answers. CQA also takes advantage of *Wisdom of the Crowd*, or the idea that everyone knows something [30]. Users can contribute to the community by asking questions, giving answers, and voting for high-quality posts.

Many initial CQA venues, such as Yahoo! Answers and AnswerBag, were developed to support general-purpose questions. Other CQA platforms focus on more specific topics; for example, Stack Overflow supports issues related to computer programming. Recently, CQA has evolved to support online learning. Some small-scale CQAs were introduced in order to support small groups of university students [2], [28]. These educational CQA sites include Chegg¹, Piazza², and Brainly³. Brainly, for instance, specializes in online learning for primary and secondary students, helping them interact with each other by asking and answering questions related to school subjects (e.g., English, Mathematics, Biology, Physics, Chemistry, etc.) [7]. Although most CQAs are publicly available and any student is welcome to join for educational purposes, a majority of new CQA users may be fully or partially unaware of these sites' community norms. These norms constitute user behaviors that govern how to appropriately ask and answer a question in order to satisfy an asker's information need. Misunderstanding these norms may affect a user's ability to post appropriate and mutually beneficial answers to a CQA community. Though users who struggle to understand community norms but continue to answer questions differ from online lurkers who most likely consume content without creating it [29], they both need guidance in order to create appropriate answers and participate in healthy question-answering activities. Thus, the main research objective of the current study is to understand characteristics of those struggling CQA users and investigate a series of features that indicate their behaviors. To do so, we propose a framework to automatically identify struggling users based on their question-answering activities. We also attempt to investigate the feasibility of detecting struggling users in the early stages of their problems by using community feedback. Community feedback is the feedback from other users to the quality of answers such as other users vote for best answers or moderators delete bad quality answers. Understanding struggling users and identifying them in the early stages

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHIIR '17, March 07 - 11, 2017, Oslo, Norway

© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4677-1/17/03...\$15.00

DOI: <http://dx.doi.org/10.1145/3020165.3020181>

¹<https://chegg.com/study/qa>

²<https://piazza.com/>

³<http://brainly.com>

of their CQA activities may help create appropriate user guidelines that demonstrate how students can properly seek and share information within an online education community in order to increase or improve their knowledge.

In the current study for identifying struggling users, we attempt to examine Brainly, one of the largest CQA services specifically targeted towards education. Brainly is an online social learning network 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, Indonesia, and more.

Our specific contributions with this work are as follows.

- Conducting an empirical study centered on struggling users to see how their actions differ from user norms.
- Proposing a framework to find the struggling users based solely on their online activity.
- Proposing a framework to detect the struggling users in their early stages through their online activity and community feedback (i.e. human judgment on the quality of answers). Early detection can help us outline suitable actions to help these users.
- Examining the importance of features in detecting struggling users. These show that some early community feedback is important.

2. BACKGROUND AND RELATED WORKS

2.1 Community Question-Answering (CQA)

The emergence of the Web, and specifically Web 2.0 technologies, have changed the way that users look for information in virtual environments. Search engines allow users to quickly find information, but they do not satisfy needs that require more personalized answers. These needs led to the development of Community Question-Answering (CQA) platforms, which have become popular places for Internet users to look for more personalized answers to their questions. Some popular CQAs, such as Yahoo! Answers and Stack Overflow, attract millions of users, and CQA takes advantage of *Wisdom of the Crowd*, or the idea that everyone knows something [30]. In CQA, any user can join and become a part of the community by asking questions, giving answers, making comments, and voting for appropriate and thorough posts. In order to ensure content quality, a small number of users are promoted to moderators, and the majority of the community's activities are curated by those moderators.

The popularity and success of CQA have attracted academic research interests. Previous works have investigated user interest and motivation for participating in CQA [24, 33]. Using the popular Yahoo! Answers, Adamic et al. [1] studied the characteristics of CQA participants. The authors analyzed questions and clustered them based on their contents. The results identified diverse user types. The study revealed that a majority of users is only interested in a small number of topics, while other users have a broad interest. The work also showed that it is possible to automatically select questions' best answers via some basic features, such as a post's length and the past answers given by its author.

Moreover, as answer quality is among the most critical factors for a CQA site's success, previous work showed that it is possible to automatically examine answer quality based on users' past history [26, 31, 11]. Shah et al. [27] discovered that questions remain unanswered in CQA when they are fact-based. This work

also explored why fact-based questions often fail to attract an answer. Automatically determining a posts' quality can significantly reduce the work CQA can require of human judgment. Le et al. [17] used different group of features to evaluate the quality of educational answers automatically. Teevan et al. [32] discovered that questions' punctuation, length and scope affect the quantity, quality, and speed of answers in CQA.

2.2 CQA for Online Learning

Early-developed CQA platforms, such as Yahoo! Answers and WikiAnswers, support general-purpose activities. Later forums such as Stack Overflow center around narrower topics and attract more focused communities. Recently, CQA has been deployed in education. Some small-scale CQA tools were developed to support small groups of university students [2], [28]. Some larger educational CQAs are becoming popular; these include Chegg, Piazza, and Brainly. Brainly specializes in online learning for students through asking and answering activities in 16 main school subjects (e.g., English, Mathematics, Biology, Physics, etc.) [7].

Educational CQA not only provides a platform to support questioning and answering, but can also test a student's knowledge based on the content they generate. A student can learn from others' answers, and can also learn through their own answering activities, as helping others allows them to test whether they understand a question and its topic.

2.3 User Behavior in an Online Community

An online community's success greatly depends on its users because community content and activities are primarily based on users' activities. Thus, user behavior attracts a great deal of research interest. A community's size—especially the ratio of answered to unanswered questions—is a large component of its overall effectiveness. [34]. Dumais et al. [8] studied users' behaviors based on the log of activity and suggested suitable methods to design the system to collect search data. They also discussed the challenges of using real users' activity logs. Users make diverse contributions in CQA. Le and Shah [16] showed that a small number of users contribute heavily to their community and are crucial to its health. This observation motivated these researchers to develop a framework in which they integrated various features to detect top contributors in their early stages of site use. Contrastingly, "the lurker" is prevalent in many online communities. Gong et al. [9] profiled lurkers in an online social network. Compared to other users, lurkers maintain far fewer social connections. The work found that popular global events will break lurkers' silence. It is not clear why these users are lurkers or why active users become lurkers.

The ability to detect an untruthful contribution is also an important task [23, 31]. Pelleg et al. [23] studied truthfulness in CQA sites. This research examined whether users provide truthful information about themselves on CQA sites, and found that askers generally provide accurate personal information, even when they post sensitive questions. Tan et al. [31] proposed a new method called CQUAL to automatically predict the quality of a post in a social knowledge base. On these platforms, such as Wikipedia or Freebase, users can edit articles. These contributions might contain inaccurate information and detract from other users' experiences with these sites. Some signals that help determine truthfulness include user contribution history, the features of each subject, and user expertise. Determining question and answer quality will ultimately prove essential when analyzing questioning and answering behaviors in CQA.

All communities want to attract new members, but existing members who leave a community are a prevalent issue. Understand-

ing when and why users leave a particular community provides a overview of the community’s health [21], [19], [25]. These works used user behavior *ego-nets* to predict behavior and activity patterns. Pal et al. [22] provided an overview of user evolution in CQA, which included various users’ in-site activities and their effects. Antisocial users also indicate community health. Cheng et al. [6] examined antisocial behavior in online postings. Research shows that it is significantly different from other kinds of behavior. For example, antisocial users compose their posts differently than others, or display a more negative sentiment. Deleting these users forms a bimodal distribution which is very high or very low. User experience with online activity also attracts interest from the research community [20], [13]. Scholars showed that user engagement is complex. Furthermore, click patterns, dwell times, and keyboard actions correlate with user engagement. Mao et al. [18] predicted users’ engagement with volunteer crowd-sourcing. The study showed that the average number of tasks was the strongest predictor of future engagement with online crowd-sourcing.

In our work, we focus on investigating struggling users who want to contribute to a CQA community, but may be unable to participate in an acceptable way. To the best of our knowledge, our work is the first study in struggling users in educational CQA.

3. DATASETS AND CHARACTERIZATION OF THE DATA

Overview: In this study, we use data provided by Brainly.com. This is an online Q&A for students and educators with millions of active users. Here, we use data from two markets: the United States (US) and Poland (PL). US is an emerging market that started in 2013 while PL is a well-established market that began in 2009. Table 1 describes some characteristics of these data sets. Brainly requires high quality answers. Thus, moderators delete incorrect answers, incomplete answers, or spam posts. A small fraction of highly experienced users are promoted to be the moderators. The moderators not only have experience, but also have a history of significantly contributing to the community. In terms of assuring answer quality, moderators are able to delete wrong, poor, and spam answers with additional explanations to answerers; approve appropriate answers; warn and ban users who behave against policies; and participate in the forum to share their experiences with content moderation. There are 207 active moderators in US and 377 active moderators in PL who voluntarily curate content on a daily basis. Moderators also evaluate the majority of new material. The posts in Brainly are divided into three levels (grades): primary, secondary, and high school. Brainly also integrates social networking into the platform, as it allows users to make “friends” and exchange ideas.

Table 1: Description about data.

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

Figure 1 plots the distribution of the number of answers given by each user. 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 an active user is willing to devote their time to helping others.

4. EXAMINING STRUGGLING USERS

To understand struggling users based on their online activities, we first show that these users exist in the community and display

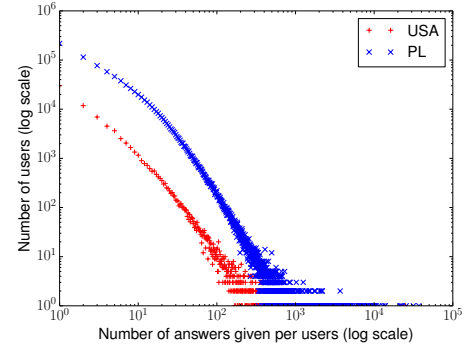


Figure 1: Distribution of number of answers given by each user. A small fraction of users answer many questions while many users answer a small number of questions.

behaviors that deviate from established community norms.

4.1 Definition of Struggling Users

In our work, we are focused on active users who have a limited ability to make meaningful contributions. We define struggling users as those who actively provide low-quality answers to other users’ questions. In particular, struggling users generate at least X posts, with a ratio of deletion of at least Y percent. The value depends on the requirement of the site. In this study, we work directly with Brainly’s data analysis and business intelligence team to determine the threshold. The Brainly data analysis team suggests $X = 10$ and $Y = 0.7$. We also extensively evaluate our method with a wide threshold range to show its efficacy. Our method works well with different thresholds.

4.2 Existence of Struggling Users

We examine the proportion of active users who are struggling within the community. We consider struggling users to be those who have the majority of their answering contributions deleted. Figure 2 plots the histogram of the deletion rate of different types of users. We plot the histogram of the deletion rates for the general population, as well as users with at least 10 and at least 20 answers. We see that some users have a higher deletion rate.

Figure 2 describes the percentage of struggling users in our data sets. We see that there are quite a few users who want to learn and contribute to the community, but cannot provide sufficient answers. These users might need significant help, making it important to detect and assist them. For example, 7.7% and 13.9% of active users in US and PL markets have a deletion rate higher than 70%. This is a large number of users equivalent to 28,364 individuals in PL’s market alone. The number of struggling users will increase when the site becomes more popular. As Brainly focuses on education, it attracts a student user population whose learning process is of particular importance.

4.3 Social Connections of Struggling Users

We examine whether there is any difference between the social connections of struggling users and those of normal users. To encourage users to exchange information, Brainly includes a social network-like structure in its architecture. We graphically represent these social connections. Each user is a node, and each friendship is an edge in the graph. The number of edges, a user’s clustering coefficient, and a user’s egonet (or friends of friends) represent some

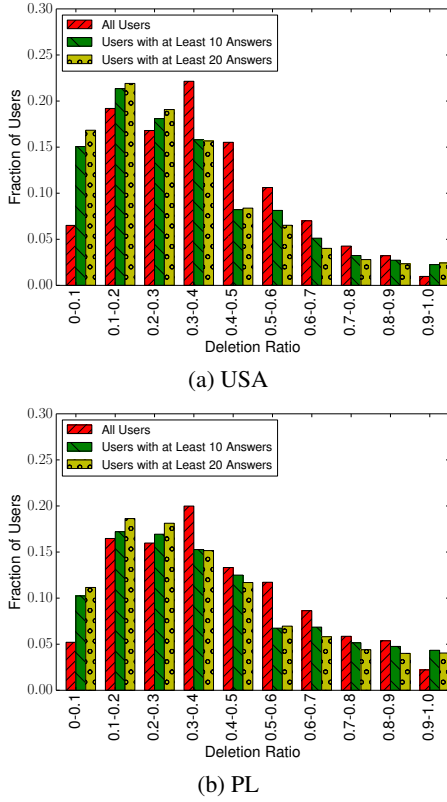


Figure 2: Deletion Rate. Histogram of deletion rate in two markets. Percentage of users with deletion rate higher than 0.7 is the sum of the last three bars (0.7-0.8, 0.8-0.9 and 0.9-1).

features that demonstrate a user’s social connections. For example, the clustering coefficient (CC_i) of a node measures how closely a user’s neighbors form a clique—or cluster together—and is defined as:

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

The higher clustering coefficient means that this user and their friends form a stronger connection. For example, a fully connected graph has (CC_i) equals to 1, and a star graph has (CC_i) equals to 0. We denote $d_i = |N(i)|$ as the number of friends of users i , while $|N(i)|$ denotes i ’s set of neighbors. The 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 use a node’s egonet features. A node’s egonet is the subgraph created by the node and its neighbors. Egonet features include its size, number of outgoing edges, and number of neighbors.

We use these features because they incorporate four social theories: Social Capital, Structural Hole, Balance, and Social Exchange [3]. The capacity of social connection plays important role in information propagation [10, 15]. Furthermore, these features can be calculated locally which is scalable for large networks.

Table 2 summarizes the social connection features of normal users and struggling users. In general, struggling users have fewer

social connections compared to normal users. Several reasons could explain this deficiency. For example, it is possible that users who provide low quality answers are less attractive to others. We perform t-tests on these two user groups. The degree and the clustering coefficient are significantly different with $p = 0.01$. A student might want to connect with someone who gives high quality answers. Thus, they are less likely to connect with struggling users. Alternatively, struggling users may not know how to enrich their social connections. In the case that struggling users do not know how to connect with good peer users, a recommendation from a new friend could help.

Table 2: Comparing the social connections between normal and struggling users. The table presents the mean values and standard error mean in the parentheses. Struggling users have fewer social connections than normal users.

Features	USA		PL	
	Normal	Struggling	Normal	Struggling
Degree	9.09 (0.41)	7.5 (1.07)	6.5 (0.08)	4.3 (0.10)
Degree_Adj	80.3 (1.60)	78.6 (1.41)	65.7 (0.3)	60.9 (0.6)
CC	0.13 (.003)	0.14 (.012)	0.04 (.002)	0.03 (.001)
CC_Adj	0.03 (.001)	0.02 (.0005)	0.09 (.002)	0.12 (.004)
Ego	25.8 (2.59)	14.2 (2.37)	14.6 (0.74)	4.5 (0.58)
Ego_Out	618.2 (19.7)	517.5 (29.6)	573.3 (7.9)	311.7 (11.2)

4.4 Reasons for Struggling Users

Fractions of deleted answers on Brainly receive comments from moderators. The site wants to enhance its users’ experiences by asking the moderators to explain why posts may be inappropriate and should be deleted. In theory, moderators’ responses will help answerers to improve their subsequent posts. Avoiding repetitive mistakes betters the entire community’s experience. Some popular reasons for deleting answers are listed in Table 3. There are many reasons answers are considered incorrect. In the majority of cases, posts are deleted due to answerers’ insufficient knowledge. In some cases, an answer might be considered low quality due to improper grammar, language, and other details.

4.5 Time to Answer Question

In the majority of tasks, greater enthusiasm and effort lead to better results. We want to see this happen in educational CQA. We measure the average time that users take to answer a question. We also measure answer length. Table 4 shows that, compared to struggling users, normal users spend much more time and effort on their answers. Additionally, answers provided by struggling users display a shorter length than those provided by the general population. We perform t-tests on normal and struggling users. The average length of answer and average time spent per answer are significantly different with $p = 0.01$. The statistics show that struggling users should put more effort into crafting high quality answers.

4.6 Difference Between Level of Education

Users with varying levels of education may not be equally capable of expressing answers. We examine whether there are differences among users from primary, secondary and high school levels. Figure 3 describes the percentage of struggling students based on their respective levels of education. The percentage of struggling primary students is higher than secondary and high school students, likely because primary students lack the ability to clearly express

Table 3: Popular reasons for deleting answers. Answers are deleted for diverse reasons. Plagiarism, for instance, indicates the answerer copied and pasted from other information sources without paraphrasing.

Reason	Example
The answer contains mistakes	Q: How do I solve $24-4*3$? A: Answer=60. $24-4=20$. $20 \times 3=60$
The answer is uninformative and does not show any work	Q: The total cost of renting a car is \$30.00 for each day, the car is rented + 28.5 c for each mile What is the total cost of renting the car for 5 days and driving 350 miles?? A: I believe the total cost is 265.5\$
The answer is pointless/meaningless	Q: When was Central Intelligence Agency founded? A: It needs to have more topic or it cant be understood or else it would be a pleasure to help you.
The answer is too vague	Q: Simplify the expression $(2x - 1)(x^2 + 3x - 2)$ A: Your answer is going to be $2x^3 + 5x^2 - 7x + 2$
The answer is incomplete	Q: How to estimate 5.804+126.009+56.2345? A: You round to the nearest thousandth
The answer contains linguistic errors or grammatical errors	Q: What color is given for desert in map? A: The answer is (BEROWN) .
The answer is plagiarized	Q: Why shouldn't we bath after having lunch or dinner? A: Following ingestion food the splanchnic circulation or blood circulation to the viscera or the different parts and organs concerned with food ...

their ideas. Fortunately, when detected and assisted at their primary (young) age, these users can significantly improve their capability.

4.7 Activeness of Struggling Users

As we mentioned, we only studied users who generated a certain amount of answers. In particular, we examined users who had at least ten posts. Among these users, we found that struggling users answered 33 questions in the US data set, and 25 questions in the PL data set. These values are lower compared to normal active users, but still high. They suggest that some struggling users want to participate in the community, but their limitations prevent them from making quality contributions.

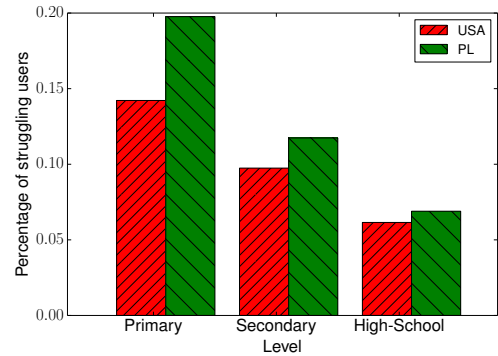


Figure 3: Percentage of struggling users with different education levels. Many primary school students struggled to provide good quality answers.

Table 4: Comparing the effort put into creating an answer. The table presents the mean values and standard error mean in the parentheses. Struggling users spent less time and effort when providing an answer compared to normal users.

Features	USA		PL	
	Normal	Struggling	Normal	Struggling
Avg Time (sec)	80.4 (0.8)	42.2 (1.3)	171.4 (0.3)	79.5 (0.4)
Avg Length	119.9 (0.9)	94.8 (1.2)	367.7 (0.6)	321.1 (1.2)

4.8 Readability of Answers

To determine the quality of struggling users' written text, we measure its readability based on two popular indexes: automated readability index (ARI), and Flesch reading ease score of answer (FRES) [12]. The ARI measures what grade level should understand the text, which is measured by

$$\alpha * \frac{\# \text{ of characters}}{\# \text{ of words}} + \beta * \frac{\# \text{ of words}}{\# \text{ of sentences}} - \gamma \quad (3)$$

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. FRES index is calculated as

$$\alpha' - \beta' * \frac{\# \text{ of words}}{\# \text{ of sentences}} - \gamma' * \frac{\# \text{ of syllables}}{\# \text{ of words}} \quad (4)$$

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.

The average FRES value for normal users is 64.1(0.4) while the FRES value of struggling users is 39.1(1.5). This reveals that it is more difficult to read the answers given by struggling users, which explains why some of their posts are deleted. But the distribution of the FRES value varies widely between struggling users. Many struggling users compose easy-to-read posts that provide incorrect information. This observation bolsters the claim that posts are deleted for a diverse array of reasons.

5. DETECTING STRUGGLING USERS WITHOUT HUMAN JUDGMENT

In the CQA community, a small number of high profile users are authorized to examine answer quality. As the community develops, users generate more content, and this places a growing burden on moderators. In this section, we examine whether it is possible to automatically detect struggling users without any human judgment. Human judgment is the actions by moderator or other users such as voting for the best answers or bad answers. The formal definition

is: “Given a user and all of their posts, can we predict, without any human judgment, whether this user has struggled with the site?”.

Finding these users is a classification problem, which includes extracting features, building the classifier and applying the model automatically.

5.1 Features Extraction

We identified a list of features based on our observations to extract struggling users. Table 5 lists the features used in our study. They are divided into four groups: *Personal Features*, *Community Features*, *Textual Features*, and *Contextual Features*. The *Personal Features* include the number of answers given by the users, users’ grade level, users’ lifetime site presence, and how much time users spend on the site. The *Community Features* include users’ social connections, as described in the previous section. The *Textual Features* include the ARI and FRES index, answers’ average length, and features to represent the format of users’ writing, such as whether their text is well formatted or whether it contains latex typing. The *Contextual Features* include the time to answer or typing speed.

These features are based on users’ activities. We excluded features generated by moderators or other human judgment. The purpose is to examine whether we can automatically evaluate the users.

Table 5: Lists of features are classified into four groups: Personal, Community, Textual, and Contextual. Features’ abbreviations are in brackets.

Personal Features	
Number of answers given (n_answers)	
Grade level of users (u_grade)	
Lifetime of users (life_time)	
Time spent with the site (spent_time)	
Community Features	
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	
The avg. length of answer (length)	
The avg. readability of answer (ari)	
The avg. Flesch Reading Ease Score of answer (fres)	
The format of answer (well_format)	
Using advance math typing: latex (contain_tex)	
Contextual features	
Duration of time taken to answer (time_to_answer)	
Typing speed (typing_speed)	

5.2 Classification

Since our problem is a typical binary classification problem, any classification method can work with our framework. In this work, we applied different popular classification algorithms, including logistic regression, support vector machine, decision trees and Random Forest [4]. Let $X = x_1, x_2, \dots, x_n$ is the list of features such as n_answers, u_grade. The list of classification algorithms is summarized as:

- Logistic regression (log-reg): Log-reg is a generalized linear model with sigmoid function: $P(Y = 1|X = \frac{1}{1+exp(-b)})$,

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

- Support vector machine (SVM) with Radial basis (RBF) kernel. The RBF kernel is defined as $K(x, x') = exp(-\frac{1}{2}||x - x'||)^2$
- 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. 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.

5.3 Experiment Setup

We perform the prediction on a balanced data set of normal and struggling users by under sampling method to create a balanced data set [5]. We perform 10-fold cross validation in all experiments. We measure the efficacy of our method with importance metrics including accuracy, F1-score, and Receiver Operating Characteristic (ROC).

- Accuracy: The accuracy measures the percentage of struggling users that we can accurately detect. The higher the accuracy, the better the algorithm.
- F1-score: This score considers both precision and recall. The precision is the fraction of instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Given precision and recall, F1-score is calculated as:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

- Area Under ROC: This involves measuring the area of the plot with true positive versus false positive rates as the threshold changes from 0 to 1. The higher the value, the better algorithm. A random guess algorithm achieves the value of 0.5, where the true positive rate increases linearly with the false positive rate.

Setting different thresholds: According to our definition, the struggling user makes at least X posts, and the ratio of deletion is at least Y percent. The value depends on the requirement of the site. Even Brainly’s analytic teams suggest $X = 10$ and $Y = 0.7$; we evaluate the efficacy of our method with different thresholds. Below is the set of thresholds tested in our method X and Y to show the efficacy of our method.

- $X = \{5, 10, 15, 20\}$
- $Y = \{0.5, 0.6, 0.7, 0.8, 0.9\}$

5.4 Results

Figure 4 plots the accuracy and F1-score when we use different number of posts (X) and deletion rate (Y) thresholds by applying Random Forest algorithm. These values show that we do not achieve high performance if we only rely on user activity and

do not include community feedback. Our method performs better with higher thresholds X and Y . Higher value X indicates that we can observe more activity and gather more information about the users. Higher threshold Y means that the users are very different from the community norm. For example, users with a deletion rate of 0.9 are very extreme users. In such cases, it is easier to detect these extreme users.

Table 6 describes the accuracy from different classifiers applied in the current study. In both data sets, Random Forest achieves the highest accuracy. Random Forest achieves the highest performance and is a scalable algorithm. In our experiment with a single machine with 2.2 GHz quad-core, 16 GB of RAM, implemented in Python code, it took less than one millisecond to classify each user. Furthermore, the tree can be built separately and is easily computed to be distributed for a larger data set.

Table 6: Comparing different classifiers. We change the thresholds of number of answers and deletion rate as in experiment setting and take the average. The table presents the average accuracy with standard deviation in the parentheses. Random Forest achieves the best accuracy.

Data sets	LogReg	SVM	Decision Trees	RF
USA	70.8 (.042)	74.2 (.028)	68.1 (.039)	73.7 (.031)
PL	75.1 (.045)	76.3 (.029)	70.7 (.037)	78.6 (.035)

We also measure the Area Under ROC curve to see the trade-off between true positive and false positive rates. The Area Under ROC for US and PL are 0.83 and 0.84, respectively. Again, these values reflect moderate performance value. We need to improve for a real application.

6. DETECTING STRUGGLING USERS AT THEIR EARLY STAGE WITH HUMAN JUDGMENT

In the previous Section, we demonstrate the low accuracy involved in predicting whether a user will struggle. In this Section, we will examine whether it is possible to predict whether users will struggle at an early stage with community feedback. When new answer is generated in CQA, other users can judge the content. Some typical human judgment includes moderator deletes the bad content or other users vote for good content. The formal definition is “Given a user with his T posts and human feedback, can we predict whether this user will struggle with the site in the future?”.

Since the moderator judges user answers, we can get some early feedback on answer quality. We use the same features as Table 5 in Section 5. We add new features, including: *del_ratio* represents the percentage of the first T answers that are deleted, and *n_best_answers* represents the number of answers that are selected as the best answer. Other features, such as readability or answer length, are similar to those in the previous section, but we only calculate the features based on the first T answers.

6.1 Experiment Set Up

We observe the community behavior and feedback on the first $X/2$ posts of each user and predict whether these users will struggle in the long term. Higher value X means we observe more activities. We also add two new features—*del_ratio* and *n_best_answers*—in the *Community Features* group. Other settings are similar to the

previous task. We also try different thresholds including the number of post X and the deletion rate Y to show the efficacy of our method.

6.2 Results

Figure 5 plots the accuracy of our method. The new features, *n_deletion* and *n_best*, significantly increase the accuracy. For brevity, we do not report the F1-score, which is similar to the accuracy. Results demonstrate that our method achieves high accuracy and F1-score in both US and PL data sets, which are both at 90% for high deletion rate threshold. These key performance metrics predict struggling users more accurately than those that measure all posts without human judgment. Furthermore, two new features significantly increase the accuracy. We will discuss the importance of each feature separately in a later part of this paper.

When using Random Forest, our method’s Area Under ROC is 0.93 and 0.94, respectively. These are very high performance metrics reiterated by Accuracy and F1-score. We see that some community feedback helps us significantly increase the system’s performance. Furthermore, we only need feedback on a user’s few initial posts to make the framework perform well, which could significantly reduce the moderator’s workload.

6.3 Experiment Results Discussion

6.3.1 Feature Importance

We measure the importance of each feature in our prediction framework. There are multiple ways to measure the importance of each feature. The general idea is to measure how a particular feature’s removal affects the framework’s overall accuracy. In the bagging method, we use a permutation test to remove the features and measure the accuracy of out-of-bag (OOB) samples. The important features will decrease the accuracy more than others.

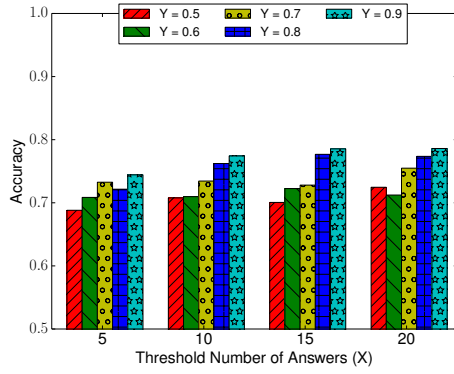
Figure 6 measures the importance of each feature in our study. In both scenarios, the average time users spend on an answer and the average length of an answer are important features. When detecting early-stage struggling users, human judgment is vital. In Figure 6b, the deletion ratio has a higher value than other features. Thus, we can achieve better performance with some human judgment in our framework.

6.4 Feature Selection

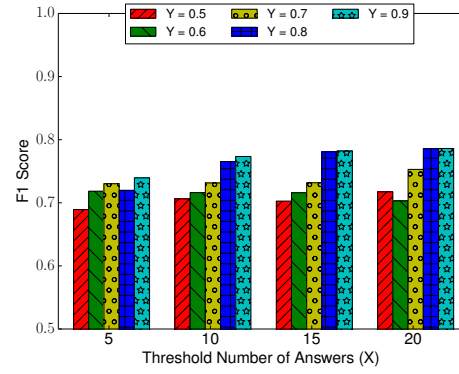
We use a small number of features in our study. Thus, feature selection does not improve performance; even some features, such as using latex, are not powerful. Furthermore, the main result presented in this work is achieved via Random Forest. When building each tree of the random forest, the algorithm already randomly selects some features. This suggests that performing feature selection is unnecessary.

6.5 Setting the Threshold

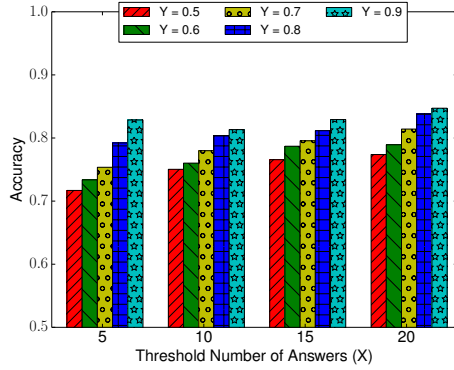
In the real system, we might choose a different threshold to identify the users who may need guidance to create appropriate answers. If we want to detect the majority of struggling users, we might need to accept a higher error rate. If we want a very low error rate, we would detect fewer struggling users. The curve in Figure 7 plots the True Positive rate against the False Positive rate. We first observe the high Area ROC in both data sets. Secondly, we can detect the majority of struggling users with the small False Positive Rate. For example, if we want to identify 80% of struggling users (True Positive Rate = 0.8), the error rate is very small. In this case, the False Positive Rate is 0.03 and 0.05, respectively. Thus, the framework is quite promising in a real system.



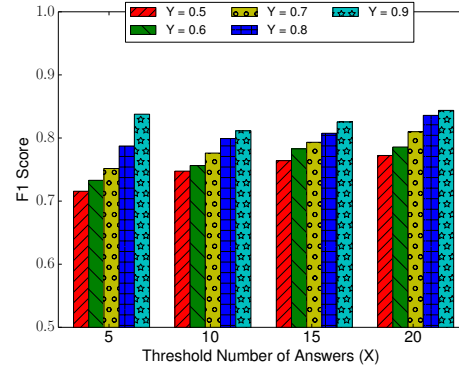
(a1) USA-Accuracy



(a1) USA-F1 Score

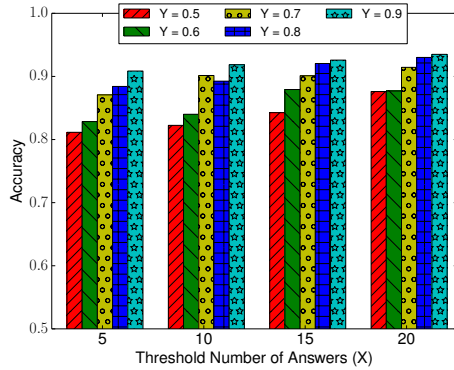


(b1) PL- Accuracy

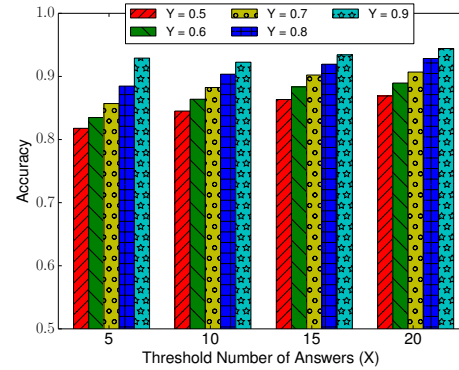


(b1) PL- F1 Score

Figure 4: Accuracy and F1 Score. The accuracy and F1 Score of detecting struggling users without human judgment. The Figures present the results when applying Random Forest with all Features. The performance is moderate without human judgment.



(a) USA



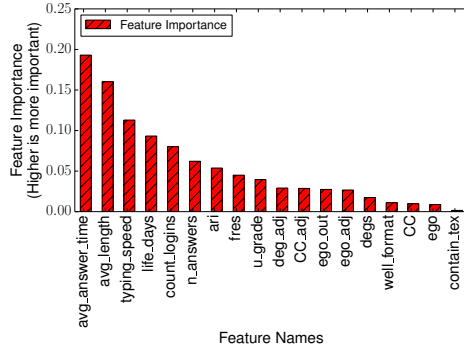
(b) PL

Figure 5: Accuracy. The accuracy of detecting struggling users in the early stage. The Figures present the results when applying Random Forest with all Features. Our method is robust to a different threshold of X and Y. Our method performs better with higher value X and Y. Users with higher deletion rate Y are different from the community norm.

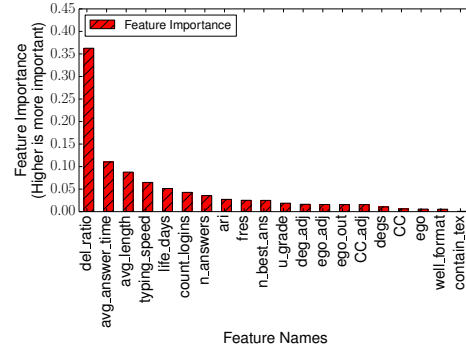
7. DISCUSSION

CQA has become an important knowledge-sharing avenue. Education is also an important field, and one in which society has always held a vested interest. Our findings show that many users are struggling within online CQA communities that are geared towards students. These struggling users should not be neglected, especially

because they hail from educational institutions. These users could be detected with high accuracy if there is some help from the CQA community. In the current system, all feedback must come from moderators, who must exert an undue amount of human effort to properly assist struggling users. Furthermore, detecting late-stage struggling users comes with a high cost in terms of those users and



(a) Without Human Judgment



(b) With Human Judgment in the Early Stage

Figure 6: Measure the importance of each feature. Average answer time and the average length of answers are important features. Furthermore, the deletion ratio in the case of human judgment is a vital feature.

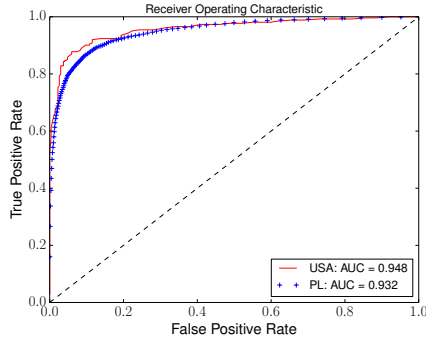


Figure 7: Plotting of ROC to select the suitable threshold. We can detect the majority of struggling users with minuscule error rate.

their wider online community; the users could be frustrated and leave the community. Furthermore, it might be too difficult to fix their knowledge gaps if they are discovered too late. Our work shows that it is possible to find struggling users in their early stage with some community feedback.

Our study has some limitations, which can be addressed in future research. First, we do not discern the reason that users are struggling in the community. Providing users with automatic, concrete feedback could improve their performance and allow them to quickly transcend their shortcomings. Second, it would be highly beneficial to directly interact with struggling users to determine how to best assist them. Observing the effect of such an early interaction would be valuable. In our future work, we hope to resolve these limitations.

8. CONCLUSION

The current work focuses on struggling users in community question answering (CQA). Some students are found to be struggling within education-oriented CQA sites, such as Brainly. These struggling users want to participate in the community, but are unable to produce acceptable answers due to their inexperience and lack of knowledge. While these characteristics may be undesirable outside of an educational context, here they are to be expected and should

be addressed to enhance students' learning experience. It follows that flagging these users simply because they are producing *bad* content may not be the best course of action. Instead, if we could use this information to identify such users and recognize that their behavior could indicate a learning-related problem and opportunity, we could help them become better users and better students. In order to identify and eventually assist these individuals, the current study compares their behavior to that of regular active CQA users. Our study reveals that struggling users want to participate in their community, but often do not possess the proper skills or attention to detail that are necessary to make valuable contributions.

Currently, in Brainly and other CQA sites, only moderators judge user performance, and this gives them an unfairly heavy workload. We show that it is possible to detect struggling users without any human judgment with moderate accuracy. However, detecting struggling users without human judgment may come too late, as many users leave the community if they are dissatisfied with their respective performances. Our framework shows that it is more accurate and effective to detect struggling users in their early stage with initial community feedback. Furthermore, using early community feedback can significantly reduce a moderator's workload.

Understanding struggling users' behavior can help us design CQA sites that better help these users. This applies to general CQA as well as educational CQA. However, detecting early-stage struggling users is particularly important in the realm of education since educators can effectively intervene and assist if they know which students are struggling. Detecting struggling users in their early stage ensures that we have enough resources and time to intervene [14]. Although the work in this paper is limited to educational CQA, the framework – including feature extraction, classifier, and evaluation – could apply to other general and topical CQA sites.

In future work, we hope to not only detect struggling users, but also to understand the more specific reasons behind their difficulties. At the moment, the CQA system still needs some input from its moderators, and automatic feedback remains the ultimate goal. To achieve automatic feedback, we might need a more complex taxonomy to analyze user-generated content. Furthermore, finding struggling users is also an important task in other domains such as search or information seeking. If we can detect struggling users in their early stage, suitable action could be taken to help them, including early intervention or content recommendation. The findings from this study give us an approach to detect such users.

9. ACKNOWLEDGEMENTS

The work reported here is in part funded by the US Institute of Museum and Library Services (IMLS) National Leadership Grant #LG-81-16-0025. A portion of this work was possible due to funds and data access provided by Brainly.

10. 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] C. Aritajati and N. H. Narayanan. Facilitating students’ collaboration and learning in a question and answer system. In *CSCW Companion*, pages 101–106, 2013.
- [3] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. Network similarity via multiple social theories. In *ASONAM*, pages 1439–1440, 2013.
- [4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.
- [5] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *The Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer, 2005.
- [6] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *ICWSM*, pages 140–149, 2015.
- [7] 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*, pages 69:1–69:4, 2015.
- [8] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*, pages 349–372. Springer New York, 2014.
- [9] W. Gong, E.-P. Lim, and F. Zhu. Characterizing silent users in social media communities. In *ICWSM*, pages 140–149, 2015.
- [10] A. Goyal, F. Bonchi, L. V. S. Lakshmanan, and S. Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social Netw. Analys. Mining*, 3(2):179–192, 2013.
- [11] H. J. Jung and M. Lease. Modeling temporal crowd work quality with limited supervision. In *HCOMP*, pages 83–91, 2015.
- [12] 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, Naval Air Station Memphis, 1975.
- [13] J. Kiseleva, K. Williams, J. Jiang, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. Understanding user satisfaction with intelligent assistants. In *CHIIR*, pages 121–130, 2016.
- [14] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *KDD*, pages 1909–1918, 2015.
- [15] L. T. Le, T. Eliassi-Rad, and H. Tong. MET: A fast algorithm for minimizing propagation in large graphs with small eigen-gaps. In *SDM*, pages 694–702, 2015.
- [16] L. T. Le and C. Shah. Retrieving rising stars in focused community question-answering. In *ACIIDS*, pages 25–36, 2016.
- [17] L. T. Le, C. Shah, and E. Choi. Evaluating the quality of educational answers in community question-answering. In *JCDL*, pages 25–36, 2016.
- [18] A. Mao, E. Kamar, and E. Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *HCOMP*, pages 103–111, 2013.
- [19] B. Ngonmang, E. Viennet, and M. Tchuente. Churn prediction in a real online social network using local community analysis. In *ASONAM*, pages 282–288, 2012.
- [20] H. L. O’Brien, L. Freund, and R. Kopak. Investigating the role of user engagement in digital reading environments. In *CHIIR*, pages 71–80, 2016.
- [21] R. J. Oentaryo, E.-P. Lim, D. Lo, F. Zhu, and P. K. Prasetyo. Collective churn prediction in social network. In *ASONAM*, pages 210–214, 2012.
- [22] A. Pal, S. Chang, and J. A. Konstan. Evolution of experts in question answering communities. In *ICWSM*, pages 274–281, 2012.
- [23] D. Pelleg, E. Yom-Tov, and Y. Maarek. Can you believe an anonymous contributor? on truthfulness in yahoo! answers. In *SOCIALCOM-PASSAT*, pages 411–420, 2012.
- [24] 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.
- [25] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, pages 732–741, 2010.
- [26] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR*, pages 411–418, 2010.
- [27] 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.
- [28] I. Srba and M. Bielikova. Askalot: Community question answering as a means for knowledge sharing in an educational organization. In *CSCW Companion*, pages 179–182, 2015.
- [29] R. P. M. L. Sun, L. Understanding lurkers in online communities: a literature review. *Computers in Human Behavior*, 38:110–117, 2014.
- [30] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.
- [31] C. H. Tan, E. Agichtein, P. Ipeirotis, and E. Gabrilovich. Trust, but verify: Predicting contribution quality for knowledge base construction and curation. In *WSDM*, pages 553–562, 2014.
- [32] J. Teevan, M. R. Morris, and K. Panovich. Factors affecting response quantity, quality, and speed for questions asked via social network status messages. In *ICWSM*, pages 600–603, 2011.
- [33] 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.
- [34] R. W. White, M. Richardson, and Y. Liu. Effects of community size and contact rate in synchronous social Q&A. In *CHI*, pages 2837–2846, 2011.