Discerning the Quality of Questions in Educational Q&A using Textual Features

Manasa Rath
School of Communication and Information
Rutgers University
manasa.rath@rutgers.edu

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

ABSTRACT

In an information seeking episode, attributes such as relevance, quality, and the nature of the information sought/obtained are directly related to the nature and the quality of the query or question that represents an information need. It is, therefore, imperative that we identify potential problems with such representation to make the information seeking outcome and the experience more successful. In this paper, we investigate the question quality for the educational community question answering (CQA) site Brainly by examining 2,000 questions, of which 1,000 were answered and 1,000 were unanswered. Two human assessors rated the quality of each question on a scale from 1-5 based on factors such as ambiguity, poor syntax, lack of information, complexity, inappropriateness, and inconsistency. We then identified different textual features that are needed to detect question quality. A logistic regression classifier was built to categorize question features based on the rating scale and textual features present in the question. The results show higher ROC curves for ambiguity, lack of information, inappropriateness, complexity and excessive information; and lower ROC values for poor syntax and inconsistency among the questions. The findings demonstrate that the classifier failed to perform when faced with ill-framed or inconsistent phrases in a question. The work described here presents a method for identifying high and low quality questions, the knowledge of which could be instrumental in helping reformulate users' questions and present them to a system or a community for more successful processes and outcomes.

Keywords

Community Q&A, question quality, machine learning, logistic regression

1. INTRODUCTION

The popularity of community question answering (CQA) sites has exponentially increased over the past several years. As a result, researchers have built an impressive body of work around CQA-related topics. These include improving users' experience when searching for answers to their information needs [13, 14] and evaluating the quality of answers given by their peers [1]. Research has also examined question askers' motivations in different contexts and situations in online Q&A [4]. Popular CQA sites include Stack Overflow, Yahoo! Answers, and Answerbag. The CQA sites for education provide an excellent platform for users to ask, learn, question and provide information using brief language. Among CQA-related research, scholars have determined why certain questions are answered over others, and what users expect when asking a question. It is now important to develop an automated process to detect and correct poorly received questions that go unanswered.

Studies demonstrate that question quality positively correlates with answer quality [1, 9]. The work in this paper describes whether automatically extracted textual features can approximately predict how humans may assess a question, and whether this approach can aid with question reformulation. In this paper, we first describe how human assessors rated questions. We then explain how we built a classifier based on the question features examined by human subjects. Finally, we discuss the implications of our classifier and how it contributes to a model for automatic question-quality detection.

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

ISBN 978-1-4503-4677-1/17/03…$15.00

DOI: http://dx.doi.org/10.1145/3020165.3022145

2. BACKGROUND

2.1 Question Quality and Types in CQA

CQA services provide a platform on which users can exchange personal information and receive social support. A wide variability exists among the types of answers askers may receive via these channels. In several cases, a question's construction can predict its answers' presence and quality. In some cases, questions asked in Q&A sites can be either informational or conversational. When asking an informational question, a user aims to receive a reply [6]. These questions are usually direct, simple, concise, and precise. An asker may pose a more complex conversational question, on the other hand, if they want to begin a discussion. However, it has also been noted that in some cases users participate to earn points due to their fascination with a site's "reward system" [11]. Belonging to different linguistic, cultural and social backgrounds may compromise perceived question quality. Therefore, on some occasions, questions may lack clarity and/or contain incoherent or awkwardly placed phrases [10]. To determine and improve question quality, various textual features such as word count, misspelled words, and sentence readability need to be taken into account.

2.2 CQA for Education and Learning

Educational Q&A sites have emerged as popular learning platforms for students. They include Chegg[1], Khan Academy[2] and Brainly.[3] In this paper we focus on a data set obtained from Brainly,

[1]https://chegg.com/study/qa
[2]https://khanacademy.org
[3]https://brainly.com

an online CQA site meant for students studying in middle school and high school. Within Brainly, students can ask questions in sixteen different subjects such as Mathematics, History and English. Brainly has spread across 35 countries and 12 languages. It is estimated that Brainly has approximately 60 million monthly users, generates around 100,000 answers a day, and has over 1,000 moderators curating its knowledge base.

2.3 Evaluating the Questions Using Machine Learning Methods

Any community member can generate a CQA site's content, meaning quality is often an issue. It is therefore necessary to automate the process of determining content quality. Existing research used various factors to detect educational answers' worth, and showed that personal features and community features are more robust in detecting valuable content [6]. Other research used askers' history, question length, and topic to predict whether a question would be answered [8]. But an unanswered question can be a good question. There are several reasons that a quality question may go unanswered. For example, other users may be uninformed on a question's topic, or a question may go unseen. In our work, we evaluate the quality of questions in educational CQA by examining a list of different criteria generated by humans. Then, we develop a list of textual features to automatically detect question quality.

3. METHOD

3.1 Defining the Problem

The central inquiry of this investigation is: given a question, can we automatically predict its quality? In this study, we classify questions based on their quality using different human-assessed criteria – such as ambiguity, lack of information, excessive information, poor syntax, etc. – all of which serve as our class labels. Textual features such as total word count, readability scores, and misspelled words help us assess the CQA questions. Here, we train the textual features along with the human ratings on the classifier to categorize each question using seven different class labels, which include ambiguity, lack of information, and poor syntax.

3.2 Obtaining Non-Textual Assessments

A total of 2,000 questions were extracted from Brainly's database in the first week of January, 2016 by writing a SQL query. From this, 1,000 questions were answered and 1,000 questions were unanswered. An equal amount of answered and unanswered questions were considered to remove sampling bias.To obtain the non-textual assessments, two trained coders were asked to evaluate the data set based on the Likert-scale items developed in [10]. The human assessors were asked to read the following explanation of various qualitative labels, then rate each question on a scale from 1-5 (least to most prevalent).

- *Ambiguity (AM)*: The manner in which the question is structured confuses the user and therefore makes it difficult to interpret the meaning of the question.
- *Lack of Information (LI):* The question does not have enough information that could be used to determine its meaning.
- *Poor Syntax (PS):* The spelling, grammar and syntax used in the writing of the question is too poor, and consequently the question is difficult to understand.
- *Complex (C):* The question is too complex, and uses difficult terminologies that cannot be answered in a particular amount of time.
- *Excessive Information (EI):* The question has excessive information that deviates from or decreases interest in the topic.
- *Inappropriate (IA):* The question uses inappropriate language, which may include taboo words, jokes, socially awkward information, and/or follow up information.
- *Inconsistent (IC):* The content of the question is not consistent with the topic, and/or the question includes (an) unrelated topic(s).

The coders were presented with and asked to rate a set of n=100 questions. It was found that the inter-coder reliability value ($\alpha >$ 0.71) was higher than the norm. Given the results, the coders were asked to code the rest of the questions.

3.3 List of Textual Features

We extracted a number of textual features that correspond to the presence of certain interrogative phrases, the number of misspellings, the total number of question marks in an entire query, the automated readability scores, the total number of characters, and the total number of words. The following list elaborates upon what these features mean and/or how they were derived.

- The presence of interrogative words such as “why," “how," “who," “where," and “what" was noted.
- We detected the presence of spelling errors to estimate the number of typos a user made. To do so, we used the Python auto correct packet.
- The presence of a question mark(s) (?) was recorded.
- The Automated Readability Index (ARI) [7] scores estimates the grade level of the user, which could comprehend the question.
- The number of characters in the question – including any special characters – were counted without the white spaces between them.
- The number of words in the question were counted to determine the question's precision or specificity.

The total number of words and characters demonstrated each question's precision. Misspelled words were detected, and suggestions for possible corrections were made. Calculated textual features suggested how each question was constructed. The readability values helped to determine the level of complexity and ambiguity present in a question's content.

3.4 Classification and Prediction Quality

With the combined results of the different textual features and the human ratings, classification of question quality was performed using the Weka v 3.9.0. Our framework works well in different classifiers like Support Vector Machine (SVM), Random Forests (RF), and Simple Logistic Regression (SLR). Here, we present the results for SLR. SLR [2] was used for ten-cross fold-validation to preserve internal validity and robustness of the model. Let $X = x_1, x_2, ..., x_n$ be the list of features. Logistic Regression (log-reg) is a generalized linear model with a sigmoid function: $P(Y = 1|X = \frac{1}{1+exp(-b)})$; where $b = w_o + \sum(w_i.x_i)$ where, w_i are the inferred parameters from regression [2]. A threshold ratings value ranging from 1-5 was set to be 3 for the individual items in every question. If the value of the rating was greater than or equal to 3, the question was labeled “Bad"; otherwise, it was labeled “Good."

Table 1: Confusion matrix for Answered Questions for Ambiguity

	Good	Bad
Good	905	14
Bad	68	13

Table 2: Confusion matrix for Unanswered Questions for Ambiguity

	Good	Bad
Good	211	13
Bad	708	68

The accuracy values were determined for all individual items, such as ambiguity, lack of information, poor syntax, complexity, excessive information, inappropriateness, and inconsistency. The values determined from those results included precision, recall, F-measure, Receiver Operating Characteristic (ROC) curves, and the class (Good or Bad) to which the question belonged. The binary classification of the question using the logistic regression classifier helped to identify a question as "Good" over "Bad".

4. FINDINGS

To evaluate the efficacy of our framework, we examine a few key metrics that include Accuracy, F1 score, and Area under the ROC curve [2]. Accuracy is the percentage of questions classified correctly. The F1 score considers both precision and recall, and can be interpreted as the weighted average of the two. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved (the recall values are not mentioned in Table 3). The F1 score values can be calculated as: $F1 = 2 * \frac{precision*recall}{precision+recall}$.

Table 3: Accuracy results found per class label.

Class Label	Precision	F-measure	ROC Curve
AM Bad	0.762	0.794	0.861
AM Good	0.863	0.833	0.861
LI Bad	0.780	0.781	0.866
LI Good	0.841	0.840	0.866
PS Bad	0.000	0.000	0.498
PS Good	0.999	0.955	0.498
C Bad	0.448	0.830	0.707
C Good	0.721	0.083	0.707
EI Bad	0.902	0.049	0.779
EI Good	0.833	0.948	0.779
IA Bad	0.578	0.281	0.723
IA Good	0.711	0.808	0.723
IC Bad	0.557	0.255	0.684
IC Good	0.683	0.788	0.684

(AM: Ambiguity, LI: Lack of Information, PS: Poor Syntax, C: Complex, EI: Excessive Information, IA: Inappropriate, IC: Inconsistent)

The results indicate significant agreement among the precision values. The most important way to determine the efficiency of the classifier is to examine the ROC curves or Area Under Curve (AUC), which indicate the number of times the classifier will correctly define an instance of "Good" or "Bad" categorization. In the ROC curve, an area of 1 indicates that accuracy is measured perfectly, whereas an area below 0.7 represents a poor performance of the classifier. The AUC values for ambiguity, lack of information, excessive information, and inappropriateness show agreeable values. However, the ROC curve values for poor syntax and incompleteness show poor results.

Table 4: Information Gain values for "Ambiguity".

Textual Features	Ranked Values
Total Characters	O.437
Total Question Marks (?)	0.01243
Interrogative Phrases	0.0124
Readability Scores	0.0075
Total Words	0.0066
Misspelled Words	0.000

The ROC curve values for ambiguity, lack of information, complexity, excessive information, and inappropriateness are decent at > 0.71. However, for poor syntax and incompleteness, the ROC values are lower, possibly because the model fails to detect question quality. This could be the auto-extraction type feature of the question that fails to recognize its special characters, grammatical errors, Internet slang, and/or informal nature. The classes we studied were highly imbalanced because the model we prepared during training is biased towards the more popular classes. This is not an uncommon scenario in data sets that have a skewed distribution of minor categories. The frequency of instances in the "poor syntax" and "inconsistent" categories was very low. As a result, our system failed to retrieve relevant instances for those categories. The precision values reported for most of the items are fairly high, showing much agreement with recall, and thereby sufficient F-measure. Usually, precision gives an idea of the fraction of relevant items retrieved. However, if the overall recall is low, it leads to low values for both true and false positives, thus leading to higher values for precision with lower ROC values. This is observed in the case of the "poor syntax" item for the "Good" class value. Similarly, we see instances where the precision values are slightly high but the ROC values are low. The chi-square statistic was found out for the good and bad questions for both answered and unanswered question. It was found that the difference between the distribution of good and bad questions between answered and unanswered questions was significant ($\chi^2 = 983.38, df = 1$,p<0.05).

Features' importance and features' selection: Considering that we only had six of the textual features present and accounted for, feature selection was not performed. However, in order to determine the amount of information gained from the textual features for a particular class label [5], the textual features were ranked using a ranking filter, and the ranked values were obtained. The results show that certain features such as total characters and the total number of question marks (?) were ranked higher compared to other textual features. The information gain values of the "Ambiguity" class label are mentioned in Table 4. Higher values indicate that a feature has more prediction value.

5. DISCUSSION

Findings based on insufficient ROC values show that the model fails to classify the questions based on poor syntax and inconsistency. Our sample is much smaller compared to the data sets used in other studies, which is a major limitation of this work. We would like to attempt to further experiment and obtain better results for the ROC curve in future research [9, 8]. On the other hand, the experiments described here show that non-textual attributes play a crucial role in determining the question quality. The study here paves a path for future research that involves developing better techniques

for parsing the misspellings and grammatical errors present in the questions. Future studies could divide the questions into categories – such as fact-based, opinion-based, and advisory – using the textual features we examined. This division would help to build specific classification models based on the textual features prominent in a particular set of questions. Future work may also include decision trees to individually detect the issues present in a particular class-label. The "poor syntax" and the "inconsistent" ROC values show that these items fail to be classified by our model. We could fix these low ROC curve values using the SMOTEBoost Methodology within the Adaboost procedure to improve the model's construction. Mining in the imbalanced datasets is a prominent problem that scientists are trying to solve [3]. The results also indicate that the unanswered questions have been significantly more (using χ^2 test, $p < 0.05$) misclassified by the classifier for "Ambiguity" than they were for the answered questions (Table 2). However, it would be interesting to see in future research how well the classifier performs in other items for unanswered questions. This demonstrates an important concern with CQA sites: that moderators of some kind should assist question-askers, as ill-framed questions that do not signify any meaning can leave the question unattended. This work also has significant implications for improved question quality, and it would be interesting to note whether answer quality improves along with questions. This would cause human intervention or an auto-completion feature to be introduced in a Q&A site, which could streamline the question-answering process and alleviate tensions within Q&A services.

6. CONCLUSION

Our study addresses an important question held by the CQA community: how can we determine question quality in (educational) CQA? This determination is important, and our study suggests that automated processes can effectively evaluate a question's quality and improve its performance. Revising and improving question quality is immensely important, as this work could facilitate immediate access to virtual learning tools. In the absence of a mechanical learning tool, a CQA site may close. Google Answers, for example, was launched in 2001 and terminated in 2006 [12]. Beyond detecting poor questions, we need a framework for reformulating them. Here, we have shown the construction of a robust model that classifies answered and unanswered questions as "Good" or "Bad." However, our model's performance could be improved in light of future research goals, including wider question reformulation and revision. Helping users form accurate questions would also increase the possibility that they would receive precise and exact answers, which is the aim of educational CQA sites.

Some previous work have involved predicting the quality of question and answers in online communities, but did not provide sufficient details. Those methods typically suggested that a question is "good" or "bad" without examining the nuance behind those labels. To the best of our knowledge, our work is the first to consider the particular factors that make a question "bad" in CQA. It determines certain features that make a question high or low quality. The results presented here demonstrate the potential benefits of giving educational CQA users more detailed feedback and suggestions.

7. ACKNOWLEDGEMENTS

The work reported in this paper is supported by the US Institute of Museum and Library Services (IMLS) grant #LG-81-16-0025. We are also grateful to Brainly for providing us with the data.

8. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194. ACM, 2008.

[2] C. M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[3] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.

[4] E. Choi and C. Shah. User motivations for asking questions in online q&a services. *Journal of the Association for Information Science and Technology*, 2015.

[5] W. B. Croft, D. Metzler, and T. Strohmann. *Search engines*. Pearson Education, 2010.

[6] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 759–768. ACM, 2009.

[7] J. P. Kincaid, R. P. Fishburne Jr, 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, DTIC Document, 1975.

[8] L. T. Le, C. Shah, and E. Choi. Evaluating the quality of educational answers in community question-answering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 129–138. ACM, 2016.

[9] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 483–490. ACM, 2008.

[10] C. Shah, M. L. Radford, L. S. Connaway, E. Choi, and V. Kitzie. How much change do you get from 40$?–analyzing and addressing failed questions on social q&a. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.

[11] I. Srba and M. Bielikova. Askalot: community question answering as a means for knowledge sharing in an educational organization. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 179–182. ACM, 2015.

[12] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341–1352. ACM, 2013.

[13] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 475–482. ACM, 2008.

[14] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.