Understanding Question Quality through Affective Aspect in Q&A Site

Jirayus Jiarpakdee Akinori Ihara Ken-ichi Matsumoto

Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, JAPAN 630-0192
tel.+81(743)-72-5318 fax.+81(743)-72-5319
{ jirayus.jiarpakdee.jc6, akinori-i, matumoto } @ is.naist.jp

ABSTRACT

Ever since the Internet has become widely available, question and answer sites have been used as a knowledge sharing service. Users ask the community about how to solve problems, hoping that there will be someone to provide a solution. However, not every question is answered. Eric Raymond claimed that how an user asks a question is important. Existing studies have presented ways to study the question quality by textual, community-based or affective features. In this paper, we investigated how affective features are related to the question quality, and we found that using affective features improves the prediction of question quality. Moreover, *Favorite Vote Count* feature has the highest influence on our prediction models.

1. INTRODUCTION

Ever since the Internet has been in wide use, learning and sharing knowledge have become easier through the World Wide Web. Notably, a question answering service called Stack Overflow, developed in 2008, has made a huge impact on how developers are finding solutions to their problems. By only querying with some keywords related to the problem, you can easily find an approach that someone has used to solve the same or a similar problem.

The number of new members on Stack Overflow is growing exponentially, resulting in thousands of questions asked daily. Of course, every question requires an answer, yet some questions are ignored and are not answered. Moreover, questions asked by new users tend to not get an answer, which in turn reduces the motivation of new users participating in the community [17].

One approach to solving the problem might be improving the way questions are written by using the Stack Overflow's guideline [1]. Also, Eric S. Raymond, who wrote "The Cathedral and the Bazaar" [16], suggested how to write questions in open community forums [2]. These guidelines could make users know how to ask a question according to the rules and regulations, which could lead to a reduction in the number of unanswered questions.

In this paper, we study the impact of affective features on the question quality. We conduct an experiment using Stack Overflow dataset provided for MSR2015 Challenge [3]. From the dataset, we extract textual, community-based and affective features(positive sentiment, negative sentiment, and politeness), and then we build models to identify the unanswered questions. Then we apply Scott-Knott test to reveal what feature is the most influential in our models. Using a sampled dataset of 38,231 questions, we have answered the following two research questions.

RQ1: ***Do affective features affect the question quality?***
From our experiment we found that using only affective features to train a model results in bad performance, but selecting affective features as one of the feature aspects to training a model improves the model's performance.

RQ2: ***How important are affective features to the question quality?***
The result from the Scott-Knott analysis showed that, one of the affective features, namely *Politeness*, is grouped in the second highest rating. Although not the highest-ranking feature, affective features are important and improve the performance of a model to identify the question quality.

This paper is laid out as follows. Section 2 introduces the background and discusses the related work. Section 3 provides the design of our experiments, and section 4 describes our case study and the results. Section 5 discusses the threats to validity of our work. Finally, Section 6 concludes this paper and describes the future work.

2. BACKGROUND & RELATED WORK

Existing studies proposed approaches to extract new knowledge from Stack Overflow [2, 10, 14, 15, 19].

Trying to understand more about Stack Overflow, Honsel et al. [10] asked a group of developers and summed up their myths about the site. From the distilled nine myths, researchers statistically checked them and found that four of them are true. They found that answered questions are

[1] http://stackoverflow.com/help/how-to-ask

[2] http://www.catb.org/esr/faqs/smart-questions.html

[3] http://2015.msrconf.org/challenge.php

SEmotion'16, May 17 2016, Austin, TX, USA

 ISBN 978-1-4503-4169-1/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2897000.2897006

asked by users with high reputation and get highly voted. Next, duplicated questions have high probability to get the answer. Last, new users tend to violate the rules more than experienced users. Their findings convinced us that community-based features have an impact on the question quality. Treude et al. [19] studied how programmers ask and answer questions, by categorizing questions into groups, labeling them, and doing statistical analysis on the Stack Overflow dataset. They found that "how-to", "review, conceptual" and "how to / novice" questions are more frequently answered.

One of our research topics of interest is the unanswered questions in Stack Overflow. Ponzanelli et al. [15] used content of a post and community-based features to create a model that can identify and filter a low quality, unanswered question. From their experimental result, their approach can reduce the size of review queues by identifying questions that are misclassified. Rather than identifying unanswered questions, Asaduzzaman et al. [2] presented an approach to predict how long does a question remain unanswered by using question's text-based and community-based features. Novielli et al. [14] proposed another approach to study the unanswered questions. They conducted a research to understand the role of emotions in Stack Overflow. They built a logistic regression model to identify successful questions that get an accepted answer using post properties, social factors and affective factors as independent variables. Their affective factors consist of sentiment and affective word classes that reveal positive or negative emotion behind the text. Our study investigates the impact of textual and community-based features to predict a question that is likely to not get an answer. Furthermore, we also focus on affective features such as sentimentality and politeness.

2.1 Sentimental Analysis

Natural language processing research commonly uses a sentimental analysis approach to identify a sentence that is positive, negative, or neutral. Thelwall et al. [18] proposed SentiStrength [4] as a measure of positive and negative sentiment from text. Positive strength scores range from "+1" (not positive) to "+5" (extremely positive). In contrary, negative strength scores range from "-1" (not negative) to -5 (extremely negative). For example, "I really love you but dislike your cold sister" contains both positive and negative sentiments in this sentence. "love" in the sentence is scored as "+3" as a positive sentiment from SentiStregth, and then emphasized by a booster word "really", that increases the score by "+1". In total, the positive strength of this sentence is "+4". On the other hand, "dislike" is scored as "-3" as a negative sentiment and is the most negative word in the sentence, so the negative strength of this sentence is "-3". Finally, this sentence sentiment strength is [+4,-3]. Without processing positive and negative sentiment analysis in parallel, some important information can be lost. The feature is also introduced by researchers that study the Stack Overflow dataset [3, 6, 9, 14, 17, 20].

2.2 Politeness

Politeness is a central force in communication, arguably as basic as the pressure to be truthful, informative, relevant, and clear [5, 8, 12]. Danescu-Niculescu-Mizil et al. [7] proposed an approach to calculate politeness of the input text. For any given text, the extracted result is one of the two possible outcomes: polite and impolite. They conducted an experiment using Wikipedia and Stack Overflow datasets and studied the relation between politeness and social power. Interestingly, they found a negative correlation between politeness and reputation on Stack Overflow, as users at the top of the reputation scale are less polite than those at the bottom.

[4]http://sentistrength.wlv.ac.uk/

Table 1: Summary of the studied data

	Data Dump	Percentage	Sampled Dataset
Questions with an accepted answer	4,596,859	57.53%	21,995
Questions - without accepted answer	2,472,706	30.94%	11,829
- with no answer	921,222	11.53%	4408
Total	7,990,787	100%	38,232

3. EXPERIMENT SETTING

To understand the impact of affective features to predict unanswered questions, we describe the measurement approach for our target features and the evaluation approach.

3.1 Dataset

We use the Stack Overflow data dump provided for MSR Challenge 2015 [21]. The data dump consists of 8 tables: badge, comment, posthistory, postlinks, posts, tags, users and votes. In our study, we randomly sampled 38,232 from 7,990,787 questions, keeping the same ratio between questions with and without accepted answer, as shown in table 1.

3.2 Definition of Good or Bad Question

Getting the right answer or solution to a problem is final goal for every user who asks questions. In Stack Overflow, users could accept an answer that is the best solution to their problem. Even if a question got many feedbacks as an answer, those answers might not be useful for users. In this study, we define a high-quality question that has an accepted answer.Novielli et al. [14] and Treude et al. [19] also have used the accepted answer to evaluate good questions.

3.3 Methodology

To investigate the impact of affective features on the question quality, our approach has two main parts, preprocessing and data analysis, shown in figure 1.

3.3.1 Preprocessing

To use the provided Stack Overflow dataset for our study, first, we need to preprocess raw data. Our preprocessing process contains two steps.

(Preprocessing : Data sampling) As a preliminary study, we reduce the size of the dataset, because it would take too much time to measure our target features from each question. In order to select a subset that will be an effective representative of all questions, we analyze a statistically representative sample of questions. To obtain proportion estimates that are within 5% bounds of the actual proportion (7,990,787 questions) with a 95% confidence level, we ran-

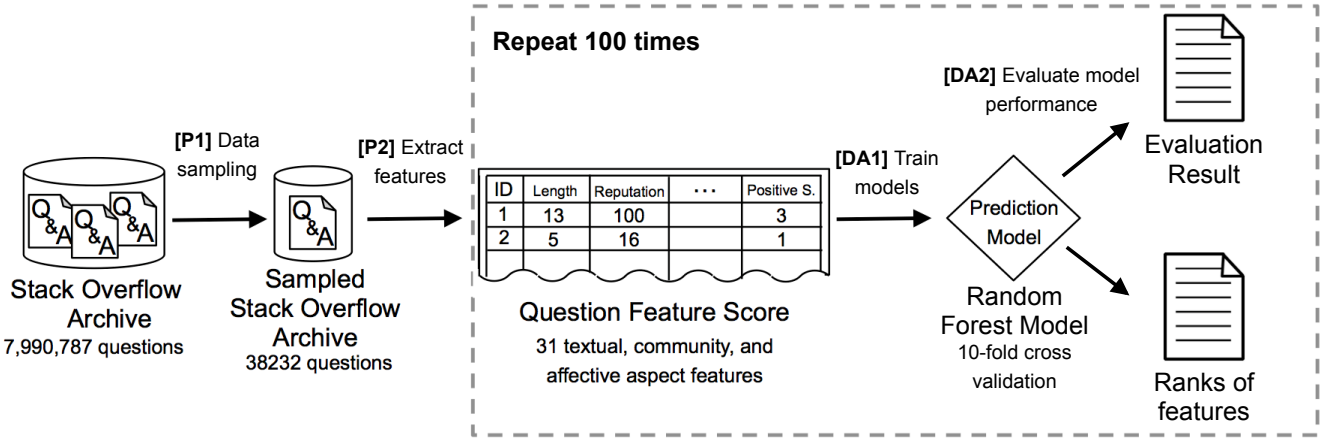

Figure 1: An overview of our preprocessing (P) and data analysis (DA).

domly select a sample of size $s = \frac{z^2 p(1-p)}{0.05^2}$, where p is the proportion that we want to estimate and $z = 1.96$. Since we did not know the proportion in advance, we use p = 0.5. We further correct for the finite population of questions P using $ss = \frac{s}{1+\frac{s-1}{P}}$ to sample for our analysis. Finally, we sampled 38232 questions. The number of answered questions and unanswered questions of sampled dataset are shown below in table 1.

(Preprocessing 2: Extract features) To train a model, we measure 31 features in terms of text-based, community-based, and affective features. Table 2 shows our target features list.

Textual Aspect. Previous study [2] found that not only "Fails to attract an expert member", but "Too short, unclear, vague or hard to follow" is also one of the top characteristics of unanswered questions. Their finding persuades us to include textual features in our study. From each question, we extract the textual features that consist of 10 features as described below. *Content Length*, *Sentence Count* and *Word Count* are directly extracted from question content after filtering out the code section. Then, we process the content to get standardized readability metrics(*Coleman Liau Index*, *Flesch-Kincaid*, *Flesch Reading Ease*, *Gunning Fox*, *SMOG* and *ARI*). Finally, for code section, we use the *Line Of Code* as it can represent how much code user put into the question.

Community-based Aspect. In Stack Overflow, badge and reputation represent how well user performed and contributed to the community. To earn a high reputation or get a badge, users not only require technical expertise but also have to regularly participate and answer as many questions as possible. On the other hand, asking a poor quality question or extremely off topic questions would likely get bad feedback and be deleted[5]. To calculate the *Reputation* of the user at question creation time, we consider votes and badges received as proposed by [15]. We also calculate *Answer Badge*, *Question Badge* and *Vote* of the user who asked the question at the question creation time. Finally, we also include *Tag Count* as one of the features.

Affective Aspect. How you write the question will affect how users perceive it. Using positive, gentle and polite statement will increase your chances of getting a useful answer, so we use sentiment and politeness of a question as features. We use SentiStrength to extract both *Positive* and *Negative Sentiment* and use the approach proposed by [7] to extract *Politeness* of a question.

3.3.2 *Data Analysis*

With the preprocessed dataset, we first train models. From those models, we then analyze their performance, and measure the influence each feature has to the question quality. Our data analysis can be divided into three steps as described below.

(Data Analysis 1: Train models) To train our models, we conduct 10-fold cross validation with random forest algorithm [4], implemented by R using bigrf package [13][6]. We repeat the process 100 times, which results in 1,000 scenarios.We conduct experiments on 7 different environment settings shown in table 3.

(Data Analysis 2: Evaluate model performance) To evaluate the model performance, we use the traditional information retrieval evaluation metrics, precision, recall and F-measure.

(Data Analysis 3: Analyze influence of features) Aiming to find the most influential features in our models, we compute variable important score using the varimp function from bigrf R package [13]. High important score shows that the feature has high influence on our models. Then, we apply Scott-Knott test [1]. The test will cluster the features according to statistically significant difference in their mean ranks. The algorithm is provided by the ScottKnott R package [11].

4. EXPERIMENT

In this section, we present the results of our experiment with respect to our research questions.

RQ1: ***Do affective features affect the question quality?***

We answer this research question with the model performance evaluation result. Table 3 shows the evaluation result of our prediction model. The result of TA (Textual fea-

[5]http://stackoverflow.com/help/deleted-questions

[6]https://cran.r-project.org/package=bigrf

Table 2: Our target features list

Aspects	Feature	Description
Textual	Content Length	The length of the question content in characters, excluding <code> section
	LOC	Line of code written in <code> section
	Sentence Count	The number of sentences in the question, excluding <code> section
	Word Count	The number of words in the question, excluding <code> section
	Automated Reading Index	$4.71 \cdot (\frac{characters}{words}) + 0.5 \cdot (\frac{words}{sentences}) - 21.43$
	Coleman Liau Index	$0.588 \cdot L - 0.296 \cdot S - 15.8$ where L is average number of letters per 100 words, S is the average number of sentences per 100 words
	Flesch Kincaid Grade Level	$0.39 \cdot \frac{totalwords}{totalsentences} + 11.8 \cdot \frac{totalsyllables}{totalwords} - 15.9$
	Flesch Reading Ease Score	$206.835 - 1.015 \cdot \frac{totalwords}{totalsentences} - 84.6 \cdot \frac{totalsyllables}{totalwords}$
	Gunning Fox Index	$0.4[\frac{words}{sentences} + 100 \cdot \frac{complexwords}{words}]$
	SMOG Grade	$1.0430 \cdot \sqrt{polysyllables \cdot \frac{30}{sentences}} + 3.1291$
Community-based	Reputation	Reputation of the user when he/she asked the question
	Answer Badge Count	The number of answer badges in total which user obtained before he/she asked the question
	Question Badge Count	The number of question badges in total which user obtained before he/she asked the question
	Tag Count	The number of tags which user assigned when he/she asked the question
	Accepted Answer Count	The number of accepted answers which user obtained before he/she asked the question
	Up Vote Count	The number of up votes which user obtained before he/she asked the question
	Down Vote Count	The number of down votes which user obtained before he/she asked the question
	Offensive Vote Count	The number of offensive votes which user obtained before he/she asked the question
	Favorite Vote Count	The number of favorite votes which user obtained before he/she asked the question
	Close Vote Count	The number of close votes which user obtained before he/she asked the question
	Reopen Vote Count	The number of reopen votes which user obtained before he/she asked the question
	Bounty Start Vote Count	The number of bounty start votes which user obtained before he/she asked the question
	Bounty Close Vote Count	The number of bounty close votes which user obtained before he/she asked the question
	Deletion Vote Count	The number of deletion votes which user obtained before he/she asked the question
	Undeletion Vote Count	The number of undeletion votes which user obtained before he/she asked the question
	Spam Vote Count	The number of spam votes which user obtained before he/she asked the question
	Moderator Review Vote Count	The number of moderator review votes which user obtained before he/she asked the question
	Approve Edit Vote Count	The number of approved edit suggestions which user obtained before he/she asked the question
Affective	Politeness	Feature which identify the politeness of the question as proposed by [7]
	Positive Sentiment	The positive sentiment score which extracted using SentiStrength
	Negative Sentiment	The negative sentiment score which extracted using SentiStrength

Table 3: Evaluation Result

Features	Precision	Recall	F-measure
T	0.426	0.355	0.387
C	0.406	0.336	0.365
A	1.000	0.032	0.061
TC	0.421	0.402	0.411
TA	0.437	0.353	0.390
CA	0.427	0.786	0.552
TCA	0.418	0.666	0.513

tures and Affective features), CA (Community-based features and Affective features) and TCA (All features) are improved compared to not using affective features. Though our results require further improvement for practical use, we found that though using only affective features results in bad performance, using them as one of the features to train a model improves the model's performance.

RQ2: ***How important are affective features to the question quality?***

To address this research question, we apply Scott-Knott test to the model with the highest F-measure score, trained by community-based and affective features. Scott-Knott then groups the features and ranks them according to their variable importance scores shown in table . The result shows that *Favorite Vote Count* is the only feature in the cluster that has the highest rank, indicating that it has the highest influence to our models.

5. DISCUSSION

In Stack Overflow, 42.47% of the questions do not have an accepted answer, from which 11.53% have no answer at all. Unanswered questions cause bad user experience as studied in [17]. Of course, our study could only extract the external features: *Content Length*, sentence or word count, and readability features. However, we cannot understand the meaning of the question content and make any recommendation of how to fix it. In particular, as presented by [2], it is not always about how you ask the question or how much detail you explain in the question, but the question itself is just "Too hard, too specific or too time consuming". We consider that to be able to completely solve this problem, further studies are required.

Table 4: Result from Scott-Knott: Ranks of features according to their variable importance

Group	Features	% of rating
1	Favorite Vote Count	8.300%
2	Answer Badge Count	7.992%
	Up Vote Count	7.991%
	Moderator Review Vote Count	7.971%
	Politeness	7.967%
	Undeletion Vote Count	7.955%
3	Tag Count	7.936%
	Negative Sentiment	7.927%
	Accepted Answer Count	7.927%
	Bounty Close Vote Count	7.925%
	Bounty Start Vote Count	7.924%
	Reopen Vote Count	7.922%
	Deletion Vote Count	7.922%
	Approve Edit Vote Count	7.905%
	Close Vote Count	7.876%
4	Question Badge Count	7.693%
	Down Vote Count	7.681%
5	Positive Sentiment	7.531%

5.1 Threats to Validity

External Validity. We sampled 38,232 from 7,990,787 question because processing the whole Stack Overflow dataset requires too much time and computational power. In addition, we conducted an experiment only on Stack Overflow dataset. While we believe that our results from Stack Overflow with large community and many questions would be common findings, the other communities such as Yahoo Answers [7] might get new findings.

To train our models, we use 31 features from three different aspects (textual, community-based and affective). Our features cover various dimensions for our model. However, we may have overlooked some other features that may also improve the performance of our models. For example, the question topics might contribute our model to improve the performance. In our future work, we would like to look for the useful features to improve our prediction model.

6. CONCLUSION

To understand the impact of various aspect features on the question quality and which feature has the highest influence, we built prediction models that predict if a question is likely to get no answer. From summarized evaluation result, we can conclude that community-based and affective features play an important role in the question quality identification. In addition, our Scott-Knott result shows that *Favorite Vote Count* from community-based features is the most influential feature. To identify the question quality, we can know beforehand if a question is likely to get an answer or not. In future, we could guide or recommend them on how to edit their question.

Acknowledgment

This work has been conducted as part of our research under the Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers.

[7]https://answers.yahoo.com/

7. REFERENCES

[1] M. K. A. J. Scott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

[2] M. Asaduzzaman, A. Mashiyat, C. Roy, and K. Schneider. Answering questions about unanswered questions of stack overflow. In *Proceedings of the International Conference on Mining Software Repositories (MSR)*, pages 97–100, 2013.

[3] B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of stackoverflow users. *IEEE International Conference on Software Maintenance (ICSM)*, pages 460–463, 2013.

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

[5] P. Brown and S. C. Levinson. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. 1978.

[6] F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli. Mining successful answers in stack overflow. In *Proceedings of the 12th Working Conference on Mining Software Repositories (MSR)*, May 2015.

[7] C. Danescu-niculescu mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 250–259, 2013.

[8] H. P. Grice, P. Cole, and J. L. Morgan. Syntax and semantics. *Logic and conversation*, 3:41–58, 1975.

[9] E. Guzman and B. Bruegge. Towards emotional awareness in software development teams. *Proceedings of the 9th Joint Meeting on Foundations of Software Engineering*, pages 671–674, 2013.

[10] V. Honsel, S. Herbold, and J. Grabowski. Intuition vs. truth: Evaluation of common myths about stackoverflow posts. In *In Proceedings of the International Conference on Mining Software Repositories (MSR)*, pages 438–441, May 2015.

[11] E. G. Jelihovschi, J. C. Faria, and I. B. Allaman. Scottknott: a package for performing the scott-knott clustering algorithm in r. *TEMA (São Carlos)*, 15(1):3–17, 2014.

[12] G. N. Leech. *Principles of pragmatics*. Number 30. Taylor & Francis, 1983.

[13] A. Lim, L. Breiman, and A. Cutler. bigrf: Big random forests: Classification and regression forests for large data sets, 2014. *URL http://cran. r-project. org/package= bigrf*.

[14] N. Novielli, F. Calefato, and F. Lanubile. Towards discovering the role of emotions in stack overflow. In *Proceedings of the 6th International Workshop on Social Software Engineering (SSE)*, SSE 2014, pages 33–36, New York, NY, USA, 2014. ACM.

[15] L. Ponzanelli, A. Mocci, A. Bacchelli, M. Lanza, and D. Fullerton. Improving low quality stack overflow post detection. In *In Proceedings of The International Conference on Software Maintenance and Evolution (ICSME)*, pages 541–544, Sept 2014.

[16] E. S. Raymond. *The Cathedral and the Bazaar:*

Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2001.

[17] R. Slag, M. de Waard, and A. Bacchelli. One-day flies on stackoverflow - why the vast majority of stackoverflow users only posts once. In *Proceedings of the 12th Working Conference on Mining Software Repositories (MSR)*, pages 458–461, May 2015.

[18] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, 2010.

[19] C. Treude, O. Barzilay, and M.-A. Storey. How do programmers ask and answer questions on the web? (nier track). In *Proceedings of the 33rd International Conference on Software Engineering (ICSE)*, pages 804–807, New York, NY, USA, 2011. ACM.

[20] 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.

[21] A. T. T. Ying. Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set. In *Proceedings of the 12th Working Conference on Mining Software Repositories (MSR)*, 2015.