Department of Statistics 2019: Predicting Community Engagement with Questions Across Online Question-Answer Fora

Candidate Number: 10140

Contents

List of Tables

List of Figures

Summary

The world wide web and the technologies that have accompanied it have given us the exceptional ability to comment on, engage with and question the world. While much attention has been given to identifying high-quality answers online, less consideration has been afforded to how we can improve our questions, which can be particularly beneficial for online question-answering communities where subject matter is often technical and expert resources are scarce.

One avenue to address issues of limited resources and information overload on online communities is to nudge questioners to enhance the "signal" of their questions before adding demand to a community, and this can be achieved by modeling and predicting positive community engagement for questions. The research presented here takes the first step towards this objective by building on and validating work already done on question quality and community engagement in online fora. By analysing question content from a diverse range of online communities, I am able to shed light on optimal thresholds for labeling positive and negative community engagement, improving upon work done in this area.

1 Introduction

The advent of the internet and the interpersonal communication technologies that have evolved from it have given us an unprecedented level of connection and potential interaction with the world. Every day, billions of individuals engage online not only with people they know, but with complete strangers from across the globe. A considerable challenge with these online interactions is widespread incivility, with substantial work being devoted to understanding and addressing this (Gervais, 2015; Berry and Taylor, 2017).

Online social question-answer (Q&A) fora present environments where community engagement (up-votes, answers, comments) and community guidelines should mitigate many of the issues experienced by other more provocative online platforms, yet these communities are not without issues of their own. Certain fora, such as popular Massive Online Open Courses (MOOCs), suffer from "information overload" where the degree of off-topic activity and discussion makes it difficult for answerers to find and engage with questions they *can* answer, let alone review all questions in the community.

Scarcity of expert resources is also a persistent problem in social Q&A systems, and thus the motivation for this research is to address precisely this imbalance by tackling question-formulation before questions place demand on expert resources. I plan to achieve this by eventually building a classification model that predicts positive community engagement with questions and provides this information to questioners so that they can be nudged into improving the "signal" of their questions. Since questions are the entry-point of every online Q&A community engagement, it is hoped that this will improve the overall functioning and development these communities.

The broad research question is therefore the following:

To what extent can we capture positive community engagement with questions on online Q&A communities?

Here, positive community engagement is defined as constructive, amicable interactions with user questions through answers, comments, votes, edits and so on. One assumption that is made is that questions are heterogeneous, i.e. they have varying levels of "quality" which evoke either positive and negative community reaction.

The research presented in this paper is but an initial step in the ultimate goal of classifying user questions and serves to build on methodologies and approaches already taken to measure question quality/community engagement. In it, I analyse a diverse range of questions in fora from the family of Q&A communities, StackExchange. I use a metric for community engagement to label questions as “good” and “bad” (receiving positive and negative community engagement respectively) and find more optimal thresholds for this labeling by calculating similarity metrics and linguistic differences across good/bad samples.

I now move onto a brief discussion of previous work in this field. This is followed by descriptions of the datasets used, pre-processing steps taken as well as exploratory analysis. I then discuss the methodology for measuring community engagement with a specifically defined variable, I present and discuss the results and lastly I make some concluding remarks.

2 Literature Review

Much work has gone into investigating online Q&A communities. Research has looked at answer quality (Jeon *et al.*, 2006; Shah and Pomerantz, 2010; Tian, Zhang and Li, 2013), behaviour of community experts (Riahi *et al.*, 2012; Sung, Lee and Lee, 2013) and question-asker satisfaction (Liu, Bian and Agichtein, 2008). Also, a common framework for engagement in Q&A communities is the optimisation of matching questions and community experts (Li and King, 2010; Li, King and Lyu, 2011; Zhou, Lyu and King, 2012; Shah *et al.*, 2018), or recommending questions in line with answerers' interests (Wu, Wang and Cheng, 2008; Qu *et al.*, 2009; Szpektor, Maarek and Pelleg, 2013).

I choose to focus on questions, not only because they have received far less attention in the literature, but because question quality impacts answer quality (Agichtein *et al.*, 2008) and because they are trivially the initial touch-point of a community/questioner interaction. It is highly likely therefore that increasing positive community engagement will improve how these communities function and evolve.

Since community engagement and question quality can be seen as two sides of the same coin ("good" questions leading to favourable community engagement), this research corresponds to a body of work on capturing question quality in online question-answer communities which I briefly discuss next. Note that while I consider community engagement a more accurate definition of what the following literature measures, I refer to "question quality" instead of community engagement to aid the discussion.

Recent work (Agichtein *et al.*, 2008; Bian *et al.*, 2009; Li *et al.*, 2012) attempted to model question quality using Yahoo! Answers, however this dataset lacks objective and definitive measures for question quality. The data that I will be using on the other hand is richer in that there a numerous proxies for question quality/community engagement available for large sets of observations. Most importantly, these variables are derived directly from the data rather than labeled manually, which enables a more objective, automatic and principled characterisation of the variable of interest.

One paper that made strides in classifying and predicting what they assume to be question quality is Ravi *et al.* (2014). Using latent topics extracted from Latent Dirichlet Allocation models on question content, they predict "question quality" with accuracy levels of 72% for the computer coding StackExchange community, StackOverflow.

Ravi *et al.* (2014) decide on using a question's `Score` as an indicator of question quality. I question this assumption and put forth the notion that a question's `Score` better characterises community engagement, since I believe it is difficult to define "quality" subjectively owing to communities valuing different facets of questions (i.e. closed-end for natural sciences or discussion-promoting in the social sciences). I thus characterise it as such and also use it as a response variable. This brings me to the aim of this paper, which is to critique and build on how to use the `Score` variable to label questions as attracting positive or negative community engagement.

3 Data

3.1 StackExchange Communities

The data I use for this analysis are question-content text from the family of online Q&A communities, StackExchange. There are more than 170 diverse StackExchange fora ranging from science-fiction world building to bicycles to quantum computing, with all the data publicly available in compressed XML files at archive.org.

I chose to use five of the largest StackExchange datasets, details of which are displayed below in table 1.

Table 1: **Dataset Details**

Forum	Questions	Answers	Description
StackOverflow	18m	27m	Q&A for professional and enthusiast programmers
Math	1.1m	1.5m	Q&A for people studying math at any level and professionals in related fields
SuperUser	415k	601k	Q&A for computer enthusiasts and power users
Russian StackOverflow	273k	310k	Q&A for programmers (Russian)
English	106k	249k	Q&A for linguists, etymologists, and serious English language enthusiasts

Source: Own calculations in Python.

For each forum, the following data is available per post in a `Posts.xml` file:

- `Id`: An identity variable for a post (chronological)
- `PostTypeId`: Indicates if a post is a question (==1) or answer (==2)
- `ParentId`: Indicates which question an answer belongs to (answers only)
- `AcceptedAnswerId`: Indicates which answer the question-asker selects as accepted (questions only)
- `CreationDate`: Indicates the date a post was originally made
- `Score`: The difference between up-votes and down-votes for a post
- `ViewCount`: The number of times a post has been viewed (not just site-registered users)
- `Body`: Main post content
- `OwnerUserId`: Indicates the user ID of a post's owner

- `LastEditorUserId`: Indicates the user ID of the last user to edit a post
- `LastEditDate`: Indicates the date a post was last edited
- `LastActivityDate`: Indicates the date that there was last activity on the post (not including views)
- `Title`: Post title (questions only)
- `Tags`: Collection of tags linked when a question is made (questions only)
- `AnswerCount`: Number of answers a question receives (questions only)
- `CommentCount`: Number of comments a post receives
- `FavoriteCount`: Number of times users favourite a question (questions only)
- `ClosedDate`: A date variable indicating if a question was closed (questions only)

This analysis will only use the `PostTypeId`, `Score`, `ViewCount` and `Body` variables.

3.2 Noteworthy elements

It is worth discussing the functioning of StackExchange sites in general to more thoroughly understand the data. Questions across fora are publicly available viewable by anyone on the internet, but posting a question on forum requires email registration with the forum. After registering, users start with 1 reputation (https://meta.stackexchange.com/questions/7237/how-does-reputation-work). The reputation levels that are key to this analysis are:

- 15: Gives you the ability to "up-vote" questions and answers
- 50: You can comment on questions and answers
- 125: You can "down-vote" questions and answers
- 2000: You can edit any question or answer.

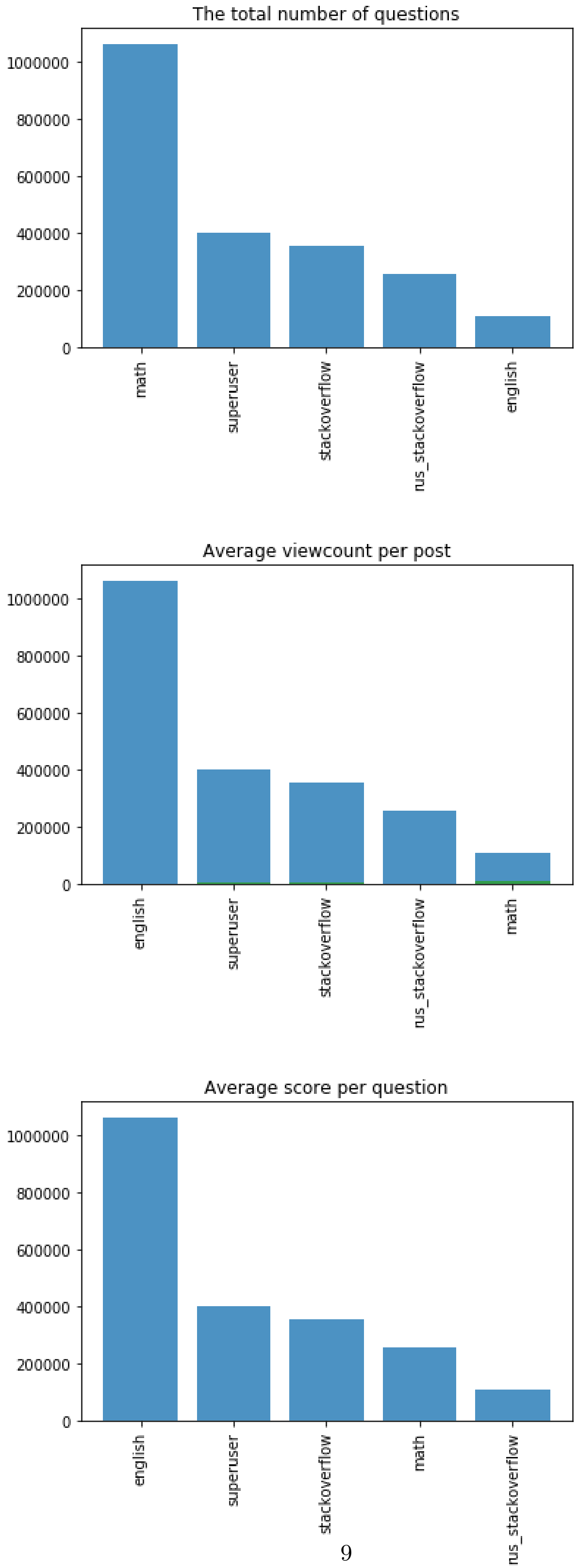
A number of methodological issues arise from how the sites operate. Firstly, owing to all questions being open to the public, many people may view questions without the ability to vote and thus still contribute to the `ViewCount` variable. Additionally, the asymmetries for privileges of up-voting and down-voting lead to a `Score` variable that is highly negatively skewed, making it appear that there are more "good" questions versus "bad" ones. Lastly, a major confounding factor is the editing of questions, not only by original posters, but also by anyone with 2000 reputation. This complicates much of the engagement between question-askers and communities because no data is available on the timing of answers, comments, votes, views etc. in relation to question edits.

3.3 Preprocessing and Exploratory Analysis

The entire analysis of the data was done with the statistical software package `R` and a link to the full code used in the analysis can be found here. After downloading and decompressing the data on the 8 selected forums, I used the `R` functions `xmlParse` and `xmlToList` from the `XML` package to parse and load the data into an `R` tibble from the `tidyverse` `R` package. Some regular expression work was needed to clean up the HTML text in questions from the `Body` variable. The `PostTypeId` variable was then used to separate out the question and answer posts, and finally the descriptive bar graphs in figure 1 were created using `ggplot2`.

In figure 1, we see that average `Score` and `ViewCount` per question vary substantially across fora. Questions on the Interpersonal and Outdoors fora have average `Scores` of approximately 17 and 11 respectively, compared to around 3 or 4 for the other fora. Interpersonal also has a significantly higher average `ViewCount` at approximately 4400 views per question followed by Fitness, Outdoors and Spanish which all have an average of around 3000.

Figure 1: **Fora Descriptive Statistics**



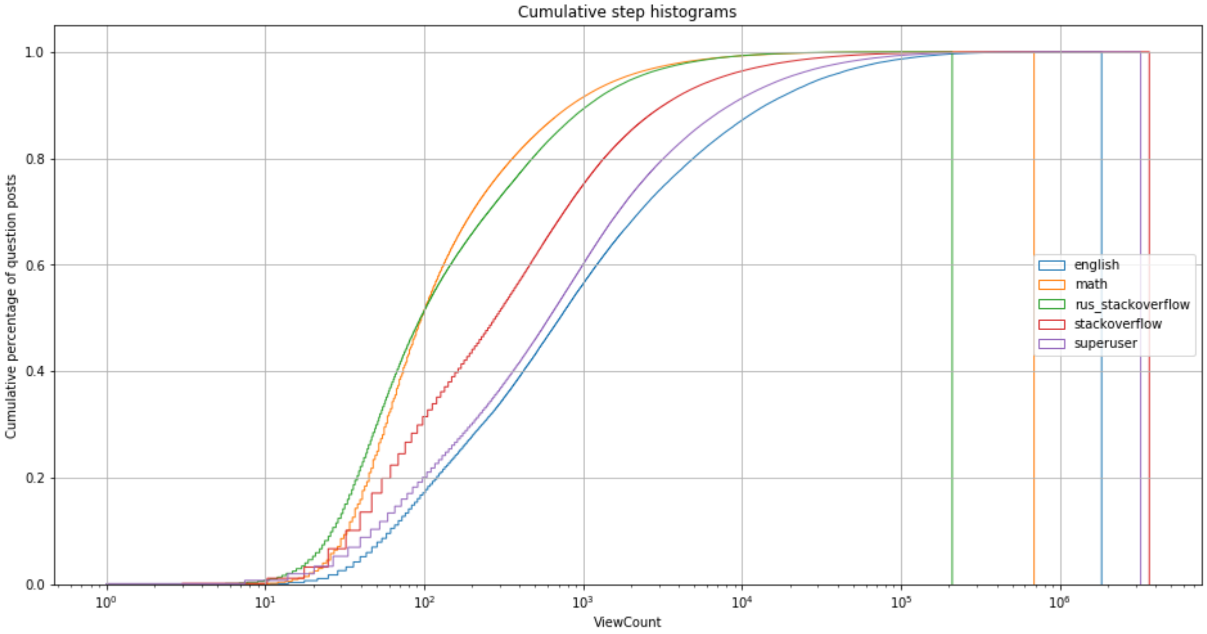
The total number of questions

Average viewcount per post

Average score per question

Source: Own calculations in Python.

Figure 2: **Cumulative Graph for Question Viewcounts**



Source: Own calculations in Python.

Figure 2 plots cumulative percentages of questions as a function of `ViewCount` in each fora. There are two aspects of the graph that are noteworthy - the order of the curves from left to right, and the fact that they maintain this order (i.e. they are roughly parallel). The further right a curve is is indicative of high `ViewCounts` per post overall, and we see that this does mirror the averages calculated in the `ViewCount` descriptive bar graph in figure 1.

A crossing of curves in figure 2 would indicate that there are certain `ViewCount` thresholds where a forum becomes more popular (experiences more viewing traffic) than another. Since there is very little crossing of curves over all fora, if at all, it appears as though the distribution of `ViewCount` across fora is fairly homogeneous and has relatively constant variance.

4 Methodology

4.1 Empirical Model

This is an equation:

$$A_{it} = f(T_i^{(t)}, S_i^{(t)}, P_i^{(t)}, B_i^{(t)}, I_i), \tag{4.1}$$

5 Results

6 Recommendations for Further Research

7 Concluding Remarks

8 References

Agichtein, E. *et al.* (2008) 'Finding high-quality content in social media', in *Proceedings of the 2008 international conference on web search and data mining.* ACM, pp. 183–194. doi: 10.1145/1341531.1341557.

Berry, G. and Taylor, S. J. (2017) 'Discussion quality diffuses in the digital public square'. Available at: http://arxiv.org/abs/1702.06677.

Bian, J. *et al.* (2009) 'Learning to recognize reliable users and content in social media with coupled mutual reinforcement', in *Proceedings of the 18th international conference on world wide web.* ACM, pp. 51–60. doi: 10.1145/1526709.1526717.

Gervais, B. T. (2015) 'Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment'. doi: 10.1080/19331681.2014.997416.

Jeon, J. *et al.* (2006) 'A framework to predict the quality of answers with non-textual features', in *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval.* ACM, pp. 228–235. doi: 10.1145/1148170.1148212.

Li, B. and King, I. (2010) 'Routing questions to appropriate answerers in community question answering services', in *Proceedings of the 19th acm international conference on information and knowledge management.* ACM, pp. 1585–1588. doi: 10.1145/1871437.1871678.

Li, B. *et al.* (2012) 'Analyzing and predicting question quality in community question answering services', in *Proceedings of the 21st international conference on world wide web.* ACM, pp. 775–782. doi: 10.1145/2187980.2188200.

Li, B., King, I. and Lyu, M. R. (2011) 'Question routing in community question answering', in *Proceedings of the 20th acm international conference on information and knowledge management.* ACM, pp. 2041–2044. doi: 10.1145/2063576.2063885.

Liu, Y., Bian, J. and Agichtein, E. (2008) '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.* ACM (Section 2), pp. 483–490. doi: 10.1145/1390334.1390417.

Qu, M. *et al.* (2009) 'Probabilistic question recommendation for question answering communities', in *Proceedings of the 18th international conference on world wide web.* ACM (2), pp. 1229–1230. doi: 10.1145/1526709.1526942.

Ravi, S. *et al.* (2014) 'Great Question! Question Quality in Community Q&A.', in *Eighth* *international aaai conference on weblogs and social media.* (1), pp. 426–435.

Riahi, F. *et al.* (2012) ‘Finding expert users in community question answering’, in *Proceedings of the 21st international conference on world wide web.* ACM, pp. 791–798. doi: 10.1145/2187980.2188202.

Shah, C. and Pomerantz, J. (2010) ‘Evaluating and predicting answer quality in community QA’, in *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval.* ACM (March 2008), pp. 411–418. doi: 10.1145/1835449.1835518.

Shah, V. *et al.* (2018) ‘Adaptive matching for expert systems with uncertain task types’, in *2017 55th annual allerton conference on communication, control, and computing (allerton).* IEEE, pp. 753–760. doi: 10.1109/ALLERTON.2017.8262814.

Sung, J., Lee, J.-g. and Lee, U. (2013) ‘Booming Up the Long Tails: Discovering Potentially Contributive Users in Community-Based Question Answering Services’, in *Seventh international aaai conference on weblogs and social media*, pp. 602–610.

Szpektor, I., Maarek, Y. and Pelleg, D. (2013) ‘When relevance is not enough: promoting diversity and freshness in personalized question recommendation’, in *Proceedings of the 22nd international conference on world wide web.* ACM, pp. 1249–1260.

Tian, Q., Zhang, P. and Li, B. (2013) ‘Towards Predicting the Best Answers in Community-Based Question-Answering Services’, in *Seventh international aaai conference on weblogs and social media*, pp. 725–728.

Wu, H., Wang, Y. and Cheng, X. (2008) ‘Incremental probabilistic latent semantic analysis for automatic question recommendation’, in *Proceedings of the 2008 acm conference on recommender systems.* ACM, p. 99. doi: 10.1145/1454008.1454026.

Zhou, T. C., Lyu, M. R. and King, I. (2012) ‘A classification-based approach to question routing in community question answering’, in *Proceedings of the 21st international conference on world wide web.* ACM, pp. 783–790. Available at: http://www2012.wwwconference.org/proceedings/companion/p783.pdf.

References