Online Community Engagement: Validating Classification Methodologies on StackExchange Fora

Bradley Carruthers, Candidate Number: 10140

3 May 2019

Abstract

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.

Contents

List of Tables

List of Figures

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 there 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 seems to be a persistent problem in social Q&A systems, and thus the rationale for this research is to eventually tackle question-formulation before they enter a community and place demand on expert resources. One approach to achieve this would be to build a classification model that can predict positive community engagement with questions, and provide this information to questioners so that they can be nudged into formulating a question that will better received by a community (improving the "signal" of their question).

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 made is that questions are heterogeneous in that they have varying levels of "quality" which evoke either positive and negative community reaction.

The research presented here 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 this paper 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 Previous work

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 the 8 largest datasets that my local machine could handle, details of which are displayed below in table 1.

Table 1: **Dataset Details**

Forum	Questions	Answers	Description
Buddhism	5.7k	19k	Discussions on Buddhist philosophy, teaching and practice
Economics	7.7k	9.9k	For those studying, teaching, researching and applying economics/econometrics
Fitness	8.2k	16k	Q&A for athletes, trainers and physical fitness professionals
Health	5.6k	4.5k	For professionals in the medical and allied health fields
Interpersonal	3.1k	13k	Q&A for anyone wanting to improve their interpersonal skills
Linguistics	7k	11k	Community for professional linguists and those interested in linguistic research
Outdoors	4.9k	12k	A forum for nature-enthusiasts
Spanish	6.4k	14k	Q&A for Spanish language linguists, teachers, students and enthusiasts

Source: Own calculations in R.

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[1]. 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.

[1] https://meta.stackexchange.com/questions/7237/how-does-reputation-work

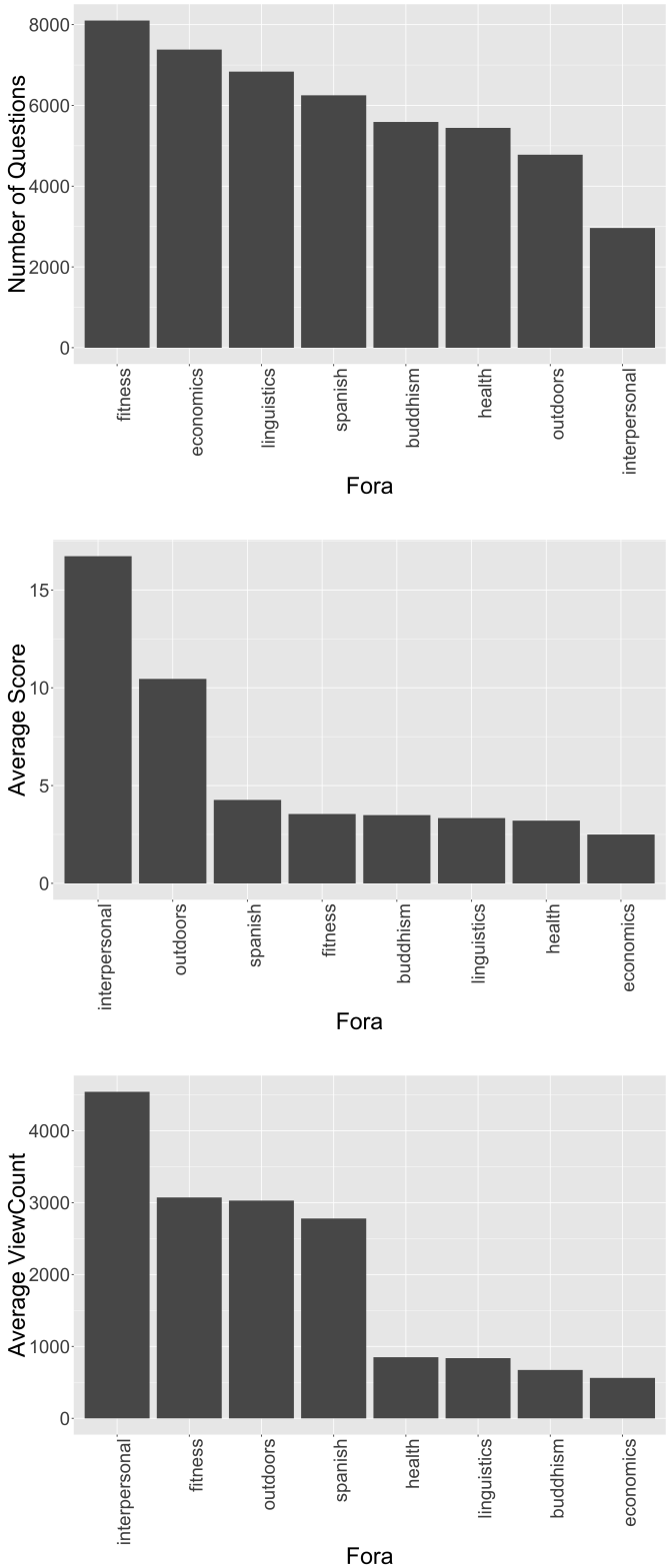
A number of methodological issues arise from how the sites operate. Firstly, owing to all questions being open to the public, many people may view questions without the ability to vote and thus still contribute to the `ViewCount` variable. Additionally, the asymmetries for privileges of up-voting and 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`.

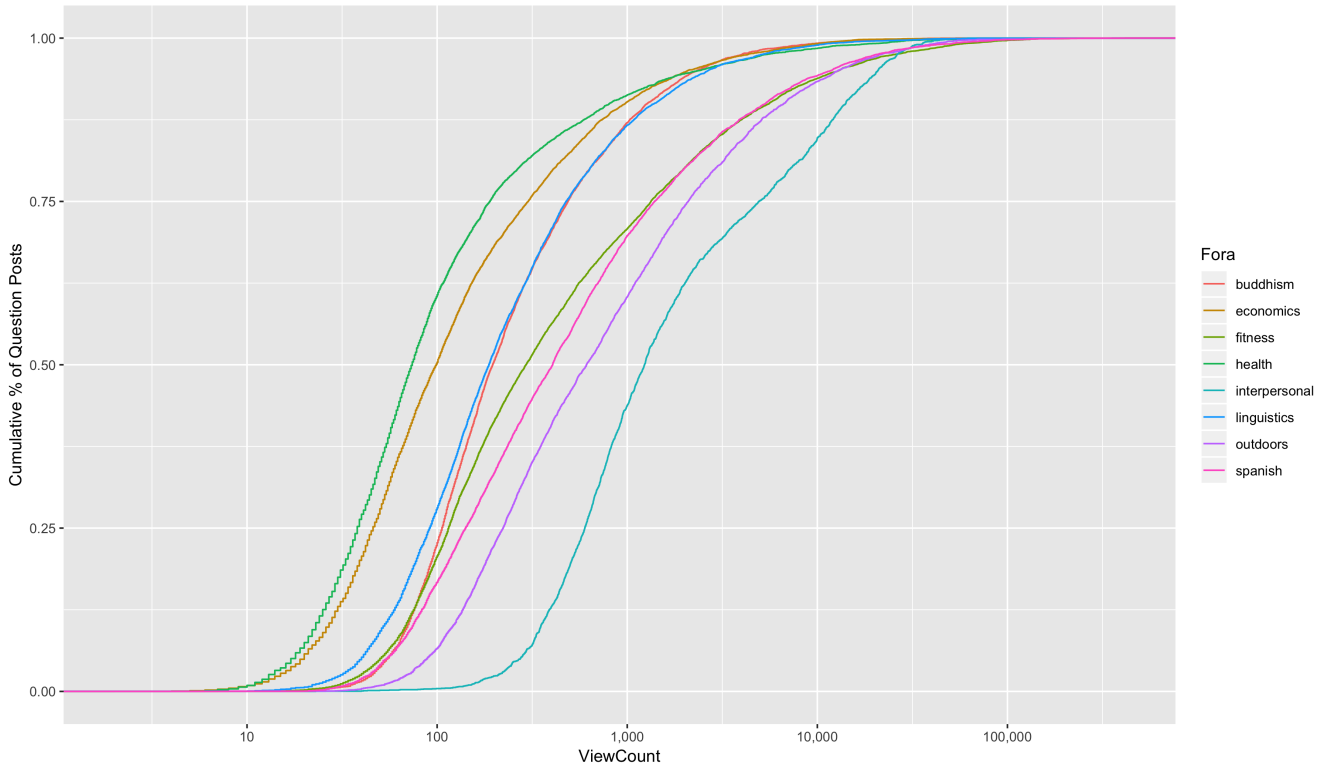
In figure 1, we see that average `Score` and `ViewCount` per question vary substantially across fora. Questions on the Interpersonal and Outdoors fora have average `Scores` of approximately 17 and 11 respectively, compared to around 3 or 4 for the other fora. Interpersonal also has a significantly higher 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**



Source: Own calculations in R.

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



Source: Own calculations in R.

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 The `Score` Variable

Now that I have parsed question content for the fora in the `Body` variable, I move onto defining a metric that measures community engagement. As previously mentioned, I choose the same metric used in Ravi *et al.* (2014), a question's `Score` to measure what I consider community engagement for questions. A point of departure here however, is that Ravi *et al.* (2014) choose to eliminate questions below a certain `ViewCount` threshold to control for the possibility that sufficient users qualified[2] to vote do not see these questions.

I feel that while it may partly be random chance that these questions do not receive views from qualified users, there may also be inherent qualities and aspects of these questions that explain this occurrence, and thus they should be incorporated into the final prediction model since this is precisely a case of a lack of community engagement. I therefore do not discard any questions from the fora.

A pertinent finding from Ravi *et al.* (2014) is that questions with higher `ViewCounts` are more likely to receive higher `Scores`. They consequently normalise `Score` by `ViewCount` to mitigate the possibility of conflating popularity with what they assume is question quality. I replicate this step in their methodology, and table 2 displays the titles of a selection of "best" and "worst" questions across fora according to the final response variable `Score/ViewCount` (the best questions have positive `Scores` whereas the worst have negative `Scores`).

Table 2 appears to show that questions that are considered the "best" tend to be honest and discussion-promoting, whereas the "worst" questions are often sarcastic and probably not genuinely looking for an answer - the questions from the Economics and Fitness fora illustrate this. Social norms also appear to play a strong part, since the "worst" question on the Interpersonal forum eludes to children being unvaccinated, which I assume would upset many individuals on the forum and lead to lowest `Score` per `ViewCount` for that forum.

[2]See section 4.2 for a discussion on registered user privileges

Table 2: **Best and Worst Fora Questions According to Score/ViewCount**

Forum	Score	ViewCount	Title
Buddhism	13	176	What are the texts that contain words which can be attributed directly to the Buddha?
Economics	-9	102	Has anyone made a successful economic prediction more than once?
Health	14	95	In which order to put on a mask, a gown and to disinfect when visiting a hospital patient?
Fitness	-5	201	What is the best way to gain size in ankle area
Interpersonal	24	1054	How can I notice if someone is speaking with sarcasm or irony?
Interpersonal	-9	1327	How can I tell if family members consider my unvaccinated kids a threat?
Spanish	20	330	Is the use of @ instead of 'a' or 'o' in order to refer to both masculine and femenine accepted?

Source: Own calculations in R.

4.2 Final Binary Response Variable

For the purpose of predicting in the classification model, Ravi *et al.* (2014) decide on a binary response i.e. labeling questions as only "good" versus "bad"[3]. This clearly ignores the fact that community engagement lies on a spectrum, but it simplifies the prediction step and thus I emulate this and leave it up to further research to predict on a range of community engagement.

The crucial issue at this stage is that Ravi *et al.* (2014) decide on a decision boundary of 0.001 for the `Score/ViewCount` response variable, above which they define good questions, and below bad questions. Other than stating that their reason for this is so that they are "confident that this reflects the good quality of [a question], rather than an incidental click on the up vote", the full motivation behind why this specific bound was chosen over others appears to be lacking.

This is not a trivial decision, since in selecting a bound one makes a strong assumption on what the ratio of good and bad questions is for a given community. This is exacerbated by the fact that the response is now binary, since questions that are on the margin and are barely labeled "good" will be treated the same as the "best" questions in the final prediction model. It is at this key step that I perform validation checks to shed light on what the optimal choice is.

[3]For clarity, when referring to good and bad questions in this research, I refer to questions being able to attract positive and negative community engagement

5 Results

The boundary of 0.001 for the `Score/ViewCount` variable used in Ravi *et al.* (2014) appears to be half of the mean for the variable - my investigations found that this results in approximately 60% of questions being labeled as "good" in the data, therefore I use this as the baseline comparison among other thresholds[4] in these results.

I compare the splitting of good/bad questions on a threshold using Cosine and Jaccard similarity[5] with the assumption being that good and bad questions will differ across these metrics owing to differences in their feature use. For each threshold tested, document feature matrices were constructed (weighted, stemmed, with no stopwords or punctuation) on both samples of good- and bad-labeled questions to calculate the similarity statistics using the `textstat_simil` function from the `quanteda` package.

For thresholds between 0% and 50% I compare symmetric samples - i.e. for a threshold of 10% corresponding to the 10% "best" questions (a high decision boundary for the `Score/ViewCount` variable) I compare the "worst" 10% of questions. In this way, questions symmetrically around the 50% mark are ignored until the 50% threshold is reached. Past the 50% mark I compare all questions labeled good with all questions labeled bad, i.e. a threshold of 60% would be compared the other 40% questions labeled bad.

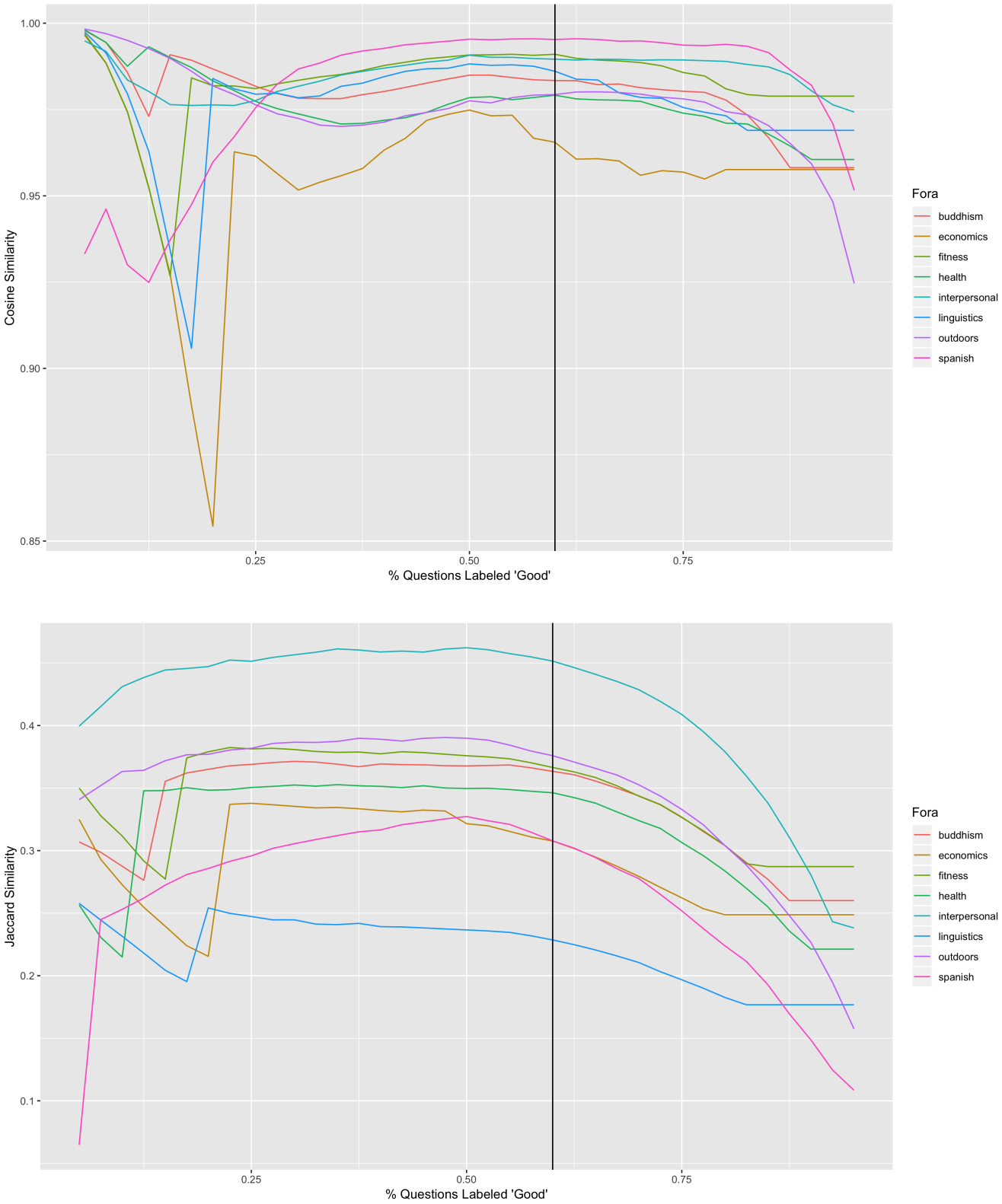
Figure 3 plots how the two similarity metrics vary over the different thresholds. Low values for these graphs indicate low similarity and thus higher distinction between good/bad samples, and the black vertical line is the 60% baseline comparison. The curves across both graphs are roughly inverted-U-shaped and similar to one another, indicating the disparity between the very "best" and "worst" questions on the far left, and on the far right showing that the very "worst" questions differ substantially from the rest of the corpus.

In both graphs on the far left we see significant oscillation of the curves - this is most likely a result of subsets of the data being compared rather than the full corpus. The low similarities of around 0.85-0.925 for the Economics, Fitness and Linguistics fora in the Cosine similarity plot are expected, since the very best and worst questions are being compared here. What is interesting is that not all of the curves in the Cosine similarity graph depict this expected behaviour, with some fora like Outdoors almost having a negative slope for the entire range of thresholds.

[4] Again in the interest of clarity, when I refer to threshold I am referring to the percentage of questions labeled "good"

[5] Euclidean distance was not included owing to it capturing essentially the same information as Cosine similarity

Figure 3: **Comparing Feature Use Across Good and Bad Questions**



Source: Own calculations in R.

Most interestingly, all curves across both graphs appear to converge towards each other and in shape from a 30% threshold onward. The fact that some curves even plateau and reach local minima after tapering off on the far right indicates not only that higher thresholds ($> 50\%$) lead to more distinct good/bad samples (which wasn't unexpected), but that an optimal threshold could be achieved where the samples are distinct and there is still a substantial number of bad-labeled questions.

Keeping in mind that the threshold decision necessitates an assumption of what the ratio of bad and good questions are in fora, I finally perform difference-of-means tests across other linguistics features for the Interpersonal and Economics[6] fora with thresholds of 80% and 75% respectively. Once again, the main assumption here is that good and bad questions differ across the features measured. The results are displayed in tables 3 and 4, with statistical significance for differences displayed for character length up to moral values[7].

Table 3: **Interpersonal Linguistic Statistical Differences**

	60% Good Questions			80% Good Questions		
Metric	Good	Bad	Diff.	Good	Bad	Diff.
N	1777.00	1185.00	592	2369.00	593.00	1776
N%	59.99	40.01	19.99	79.98	20.02	59.96
Punctuation Types	36.00	42.00	-6	41.00	36.00	5
URL-Text Types	28.00	25.00	3	49.00	0.00	49
Stopword Types	346.00	341.00	5	357.00	314.00	43
Character Length	1636.95	1774.04	-137.09**	1690.81	1695.74	-4.93
Token Length	337.01	368.06	-31.06***	348.66	352.53	-3.87
Guiraud's Root TTR	8.83	8.87	-0.04	8.89	8.68	0.21***
Yule's K	158.90	158.44	0.46	157.27	164.50	-7.23**
Flesch-Kincaid	9.35	9.14	0.2**	9.29	9.14	0.15**
Gunning's Fog Index	12.18	11.92	0.26***	12.12	11.90	0.22***
Sentiment	0.22	0.21	0.01	0.22	0.22	0
Moral Values	0.55	0.51	0.04*	0.54	0.51	0.03

Note: Wilcoxon Rank Sum Test statistical significance represented at the 0.05, 0.01 and 0.001 level by *, ** and *** respectively.

Source: Own calculations in R.

[6]The results of only two fora are presented for the sake of brevity

[7]The Wilcoxon Rank Sum Test is used owing to it being non-parametric and thus not assuming a distributional form for the data

There are a number of noteworthy results from table 3 for the Interpersonal forum. The first is that the new threshold has resulted in starker differences for some linguistic features, but not for others. This hints that our assumption of distinctions necessarily existing across all features between samples may be incorrect.

Secondly, the split for the new threshold results in a good-sample that uses far greater selections of URL-text, stopwords and punctuation. This implies that the new threshold is separating out questions that are more linguistically complex and potentially show more prior research (by linking URLs in the question `Body`). This is confirmed by the larger differences of lexical complexity (Guiraud's Root TTR and Yule's K) for the new threshold, now also both statistically significant at the 0.001 level. Lastly, the difference in character/word lengths and appeal-to-moral-values of good/bad questions has diminished for the new threshold, whereas the results for readability (Flesch-Kincaid and Gunning's Fog Index) and sentiment are essentially the same for both thresholds.

Table 4: **Economics Linguistic Statistical Differences**

	60% Good Questions			75% Good Questions		
Metric	Good	Bad	Diff.	Good	Bad	Diff.
N	4428.00	2952.00	1476	5535.00	1845.00	3690
N%	60.00	40.00	20	75.00	25.00	50
Punctuation Types	86.00	86.00	0	94.00	73.00	21
URL-Text Types	577.00	390.00	187	707.00	243.00	464
Stopword Types	346.00	334.00	12	356.00	315.00	41
Character Length	841.03	705.12	135.91***	835.84	639.16	196.67***
Token Length	189.00	153.83	35.17***	187.59	136.98	50.61***
Guiraud's Root TTR	6.71	6.29	0.42***	6.67	6.16	0.52***
Yule's K	261.48	299.75	-38.27***	264.80	312.76	-47.97***
Flesch-Kincaid	11.58	10.95	0.63***	11.50	10.81	0.7***
Gunning's Fog Index	14.86	14.25	0.61***	14.79	14.09	0.7***
Sentiment	0.24	0.22	0.02*	0.24	0.22	0.03**
Moral Values	0.48	0.44	0.04***	0.47	0.45	0.03**

Note: Wilcoxon Rank Sum Test statistical significance represented at the 0.05, 0.01 and 0.001 level by *, ** and *** respectively.

Source: Own calculations in R.

For the results of the Economics forum in table 4, there are in fact greater differences across all metrics with approximately the same statistical significance results. This is substantial evidence that a threshold of 75% results in higher discernment between good and bad question samples, while still maintaining at least a quarter of questions labeled as bad. Overall, the results above indicate that unique and optimal thresholds can be investigated and implemented across all fora, providing more robust binary samples for prediction and classification of community engagement.

6 Conclusion

The discussion and results of this paper demonstrate that the decision on where to label good and bad questions for online question-answer communities using the `Score/ViewCount` variable, or any response variable for that matter, should not be taken lightly. By analysing the distinguishing linguistic factors of various thresholds of good-labeled questions, I was able to establish that more optimal decision boundaries can be obtained for this variable in the interest of more robustly capturing and predicting positive community engagement.

It is recommended that future research analyse further Q&A community data and particularly the StackOverflow forum used in Ravi *et al.* (2014) in order to ascertain how their threshold compares to potentially more optimal values. Further work could also attempt to capture and predict on a range of community engagement rather than the discrete binary response defined in this paper.

One area of further research that I plan to pursue is to take these results and implement them in a binary classification model of my own, with the aim of predicting questions that elicit positive community engagement. In this way, I hope to assist online Q&A fora users in improving their questions, and consequently improve the prosperity of online Q&A communities as a whole.

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.