Department of Statistics 2019

Predicting Community Engagement with Questions Across Online Question-Answer Fora

Candidate Number: 10140

Submitted for the Master of Science, London School of Economics, University of London

August 2019

Table of Contents

List of Figures

List of Tables

Summary

Formulating constructive questions and receiving answers to these questions is not only a key part of how we learn as humankind, but of scientific progress and development. The evolution of the world wide web and the technologies it has provided us has given us an unprecendented ability to engage with and learn from the world, and while substantial attention has been dedicated to finding correct answers (just ask Google), comparitively less has been devoted to how we can improve the constructiveness of our questions. One online setting where relevant and well-researched questions is of particular importance is online question-answer (Q&A) communities where expert resources are scarce in comparison to cascades of new questions, or what has been termed information overload. This research aims to address this problem by analysing a diverse range of StackExchange Q&A communities: I build on work concerning questions in online Q&A communities and construct models to predict on the community-granted score for each question as a measurement of community engagement, using only textual question content available at the time questions are initially submitted. I confirm findings of prior research showing that features derived from the length and topics of questions yield improvements over baseline predictions, and also am able to gain insight into the degree of heterogeneity across online Q&A communities. My analysis shows that there is still much to be done to predict online community engagement effectively, especially on data with temporal aspects. Nevertheless, this research serves as a stepping stone to accurately informing questioners of how their questions will be received by an online community and potentially nudging them to improve their questions before adding demand to these communities, thereby improving the functioning and efficiency of these platforms substantially.

1 Introduction

Modern interpersonal communication technologies made possible by the internet have afforded us an exceptional level of connection and engagement with the world. Billions of individuals now interact online instantly, not only with people that they know, but with strangers millions of miles away. Online question-and-answer (Q&A) websites such as Yahoo! Answers, Quora, the StackExchange family and forums of Massive Online Open Courses (MOOCs) have become an extremely popular way in which internet users share knowledge about specific fields and subjects. These websites serve as platforms where users seek answers to and discussions on **complex and technical** questions that modern search engines are evidently yet unable to fully address.

It goes without saying that producing legible, relevant and well-researched questions in online Q&A fora is particularly valuable, not least since platforms are particularly prone to "information overload" where cascades of new questions far outweigh the few expert resources available. This research aims to address this problem by determining to what extent positive community engagement can be predicted using only the textual content of questions - i.e. a question's `Title` and `Body`.

The broad research question can therefore be defined as the following:

To what extent can community engagement with questions in online Q&A communities be accurately predicted?

While there is a substantial amount of literature that has addressed Q&A fora, it has focused on identifying expert users and high quality answers rather than given attention to questions, despite questions being the entry point for every interaction in communities. I draw heavily on and critique prior research done by Ravi *et al.* (2014) and analyse question content from the StackExchange family of Q&A communities. These Q&A fora have a voting mechanism whereby registered community-members can signal how much value specific questions add to the community and it is precisely this metric which I identify as community engagement and aim to predict on.

Community engagement

This research goal thus takes the form of quantitative prediction task rather than qualitative, causal or inferential analysis. I leave it to further research to address the *how* and *why* of community engagement on online Q&A communities, rather than just the *what* that is explored here. I build a **elastic net, regularised** regression model to predict the community-assigned `Score` for each question, which is the result of aggregating all community up-votes and down-votes. I evaluate models using root-mean-square error (RMSE).

This research thus has a concrete application to the real world: providing these predictions to questioners in real-time can encourage them to improve the "signal" of their questions before they submit new questions and add demand to the resources of a community. **In this way, it is hoped that the functioning and evolution of these online communities can be improved.** I believe that the fact that this research aims to predict a community-provided measurement of online engagement, has a direct real-life application and will be implemented on **diverse communities makes it the first of its kind.**

I find that models including features derived from the lengths and topics of questions perform slightly better than a baseline of just mean `Score` prediction, indicating that **there is still much work to be done and more features to be engineered before we can accurately predict community engagement in online Q&A fora.** I also find that different models various levels of performance across fora, contradicting previous research claims that topic models would be universally successful across disciplines and fora. Lastly, contrary to prior research I evaluate the best performing model using a temporal train/test split, i.e. where the training set contains questions that chronologically precede the testing set questions. Here I find **almost no gain in predictive performance**, leading me to believe that **models lacking a temporal element will always underperform when considering the chronological order of questions**.

Previous work in the field of questions in online Q&A fora will now be discussed in more detail. This will be followed by a discussion of the data, exploration of the variables included in the model, a description of the model used, and a presentation and discussion of the results. Finally, some recommendations for areas of further research and concluding remarks are made.

2 Literature Review

2.1 Question-Answer Communities

There is a substantial collection of research that has investigated Q&A communities. This includes work on answer quality (Jeon *et al.*, 2006; Shah and Pomerantz, 2010; Tian, Zhang and Li, 2013), satisfaction of questioners (Liu, Bian and Agichtein, 2008) and the behaviour of highly productive, "expert" community members (Riahi *et al.*, 2012; Sung, Lee and Lee, 2013). Two common frameworks for prior work has been the optimisation of routing questions to experts (Li and King, 2010; Li, King and Lyu, 2011; Zhou, Lyu and King, 2012; Shah *et al.*, 2018), and matching questions in accordance with answerer interest as a recommendation system (Wu, Wang and Cheng, 2008; Qu *et al.*, 2009; Szpektor, Maarek and Pelleg, 2013).

The framework I use for this research is discordant of a systems-based optimisation of question-answer matching, and is instead placed in the framework of how communities engage and react to phenomena. I focus on questions rather than user or answer characteristics, not only because they have received substantially less attention in the literature, but it has been shown that the quality of questions can significantly impact the quality of answers (Agichtein *et al.*, 2008).

With the promise of a real-life application in nudging users to enhance their question content before encumbering community resources, I believe that the development and functioning of these communities can be substantially improved if community engagement in online Q&A fora can be successfully predicted. Owing to a large overlap between literature on question quality in online Q&A fora and what I have defined as community engagement, I discuss this literature next.

2.2 Question Quality

LOTS OF WORK

High-quality questions assuredly lead to positive community engagement, however **the only difference may be the specific aspects of question content that communities value across communities.** Thus, while the literature discussed here refer to measuring and predicting "question quality", I assert that "community engagement" is a more robust interpretation of

what they are measuring and so for the sake of discussion I will refer to question quality as well.

Recent work has looking at predicting question quality for the large Q&A community Yahoo! Answers (Agichtein *et al.*, 2008; Bian *et al.*, 2009; Li *et al.*, 2012), but while this dataset has metrics for assessing answer quality in the form of answer "up-votes", it lacks a similarly community-attributed and objective metric for question quality. Agichtein *et al.* (2008) thus define question quality using question semantic features (lexical complexity, punctuation, typos etc.), Bian *et al.* (2009) manually label 250 questions and Li *et al.* (2012) combine the number of answers, number of tags, time until the first answer, author judgement and domain expertise to construct their ground truth.

Fortunately, my datasets are from the StackExchange family of Q&A fora which are rich in community engagement variables like question up-/down-votes and view-counts. Coming directly from the data, these metrics are objective rather than human-labelled and are also therefore not limited in terms of samples from the data (we can use the whole dataset).

The predictive models employed in the question quality literature have also evolved substantially. Previous work has modelled question quality based on the reputation of the questioner, question categories and lexical characteristics of questions (length, misspelling, words per sentence etc.) (Agichtein *et al.*, 2008; Bian *et al.*, 2009; Anderson *et al.*, 2012; Li *et al.*, 2012).

A fundamental distinction is that I use only the features available at the time a question is initially asked which is congruent with the goal of being able to provide real-time information to questioners before they submit questions to a community. I also don't use any features derived from user attributes, since doing otherwise would not work well for questions asked by new users.

2.3 Community engagement?

2.4 Temporality?

2.5 Ravi *et al.* (2014)

LOTS OF WORK

A paper that made much headway in the classification and prediction of what they assume

is question quality is Ravi *et al.* (2014), who use the largest and oldest StackExchange site, StackOverflow I mirror much of the analysis in Ravi *et al.* (2014), however I believe I build and diverge from their analysis significantly in a number of ways.

As discussed, much of the literature ir oriented towards "question quality" and Ravi *et al.* (2014) decide to incorporate a question's `Score` into their ground truth for question quality, yet I posit that what these studies are measuring is instead more accurately characterised as community engagement. My opinion is that question quality is much more nuanced than prior research has asserted, i.e. while most communities will value universal aspects of questions like legibility, coherence, relevance and prior-research, it is difficult to accurately define how much of this contributes to a universal inherent "quality" objectively compared to community-specific traits that communities will naturally value (i.e. closed-end questions in the natural sciences, discussion-promoting for social sciences). Thus while I also incorporate the `Score` metric into a response variable, my characterisation of this ground truth as community engagement is broader and more inclusive.

Another departure from the analysis in Ravi *et al.* (2014) that I make, is I consider a far more diverse range of communities to compare how models perform across fora. Ravi *et al.* (2014) specifically state that "[their] methods do not rely on domain-specific knowledge" and therefore "[they] believe [the methods] are applicable to other CQA settings as well" - this is something I particularly want to test and believe will be interesting to flesh out.

A last distinction between my analysis and Ravi *et al.* (2014) is that they treat the research aim as a classification problem, quite arbitrarily defining a threshold for their response variable to distinguish between "good" and "bad" questions. Despite **making it a more complex problem**, I opt to predict on a continuous response since that would provide a better indication to users of how well it is predicted that a community will react to their question.

Ravi *et al.* (2014) manage impressive results however: using textual features and latent topics extracted from question content (i.e. question `Title` and `Body`), their predictions on `Score/ViewCount` yield accuracy levels of 72% on their StackOverflow dataset. I will be emulating this part of their research, and thus a discussion of topic modelling is necessary.

2.6 Topic Modeling

LOTS OF WORK

Bayesian models have recently achieved immense popularity to solve a diverse range of structured prediction challenges in Natural Language Processing (NLP) (Chiang *et al.*, 2010). Blei, Ng and Jordan (2003) presented Latent Dirichlet Allocation (LDA) topic models as generative Bayesian models for documents to uncover hidden topics as probability distributions over words. LDA can therefore be useful in unearthing underlying semantic structures of documents and to infer topics of the documents.

Attaining accuracy scores of up to 72% for "good" and "bad" questions, Ravi *et al.* (2014) have indeed shown the capability of using latent topics derived from LDA modeling. Ravi *et al.* (2014)'s final predictive model is based on work by Allamanis and Sutton (2013), who also analysed the StackOverflow dataset, but did not look at **any form** of "question quality". They uses LDA models at three levels: across the whole question body, on code chunks in the question body, and on the question body without noun phrases.

Ravi *et al.* (2014) choose to model latent topics 1) Globally in order to capture topics over questions as a whole, 2) locally to seize sentence-level topics, and finally use a Mallows model (Fligner and Verducci, 1986) for a global topic structure to administer structural constraints on sentence topics in all questions.

Since Ravi *et al.* (2014) see no substantial gains in predictive accuracy using the Mallows model, I only employ the LDA features (**and maybe word-embedding features**), I also split my train/test temporally and differ in deleting low viewcount questions. Despite mirroring the methodology in Ravi *et al.* (2014), I critique and build on it in many ways and will begin this discussion on my methodology now.

3 Methodology

3.1 Data

The StackExchange family of online Q&A fora are a diverse range of over 170 community websites covering topics from vegetarianism to quantum computing to bicycles. Over and above the textual content of all questions, answers and comments posted since each communities conception, rich meta-data on all communities is publicly available in XML files compressed in 7-Zip format at archive.org.

The **five** datasets that I chose to analyse are displayed in table 1, along with a short description.

Table 1: **Details of Datasets**

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
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, English language enthusiasts
Fitness	8.2k	16k	Q&A for athletes, trainers and physical fitness professionals
Economics	7.8k	9.9k	For those studying, teaching, researching and applying economics/econometrics
Buddhism	5.8k	19k	Discussions on Buddhist philosophy, teaching and practice
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

The entire PySpark codebase for the processing and modelling of the data done can be found here. I downloaded the data from the 5 selected fora, decompressed the 7-Zip files, converted the XML files into Parquet format and extracted the full number of questions from each forum for analysis, resulting in a total number of questions of 30 636 across all fora. From the data, the following variables are of interest to my analysis:

- `Score`: The difference between registered-user attributed up-votes and down-votes for a question
- `ViewCount`: A counter for the number of page views the question receives (both registered

and non-registered users views are taken into account)

- `Title`: The text of the question title
- `Body`: The text of the question body
- `CreationDate`: A datetime variable indicating when the question was initially posted

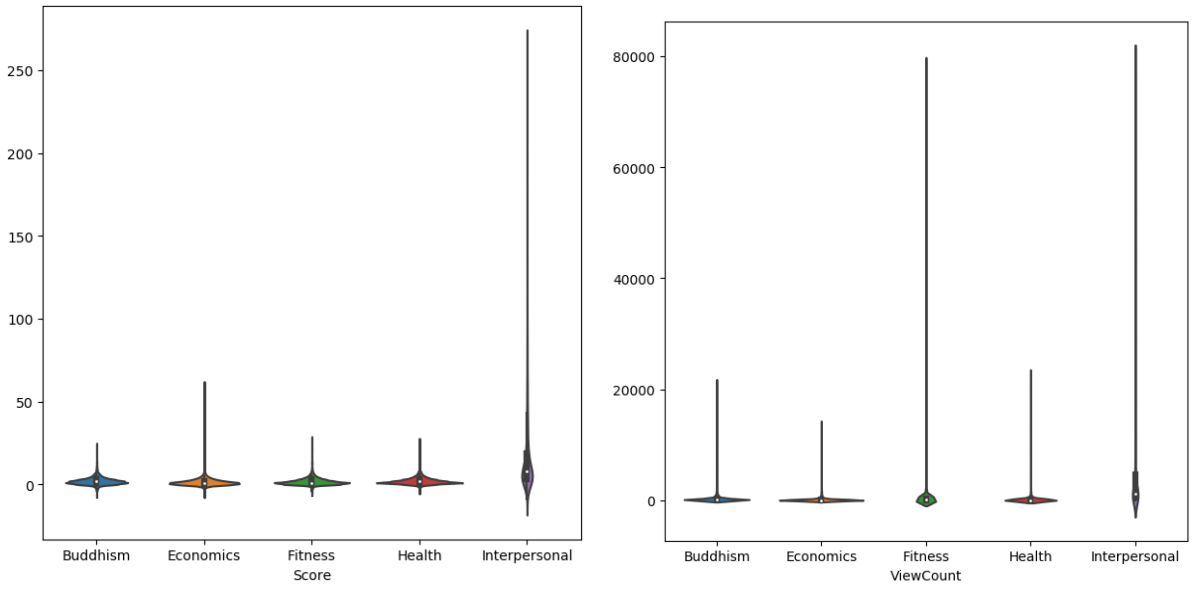
I use the most recent 3120 questions from each dataset. This evidently means that the time period that each forum's dataset ranges over will be different, potentially opening the analysis open up to bias in the form of different temporal effects across the different time periods for each dataset, but since this analysis will not be a temporal one, I assume there are no confounding temporal effects.

I now move onto an exploratory analysis of the data.

3.2 Exploratory Analysis

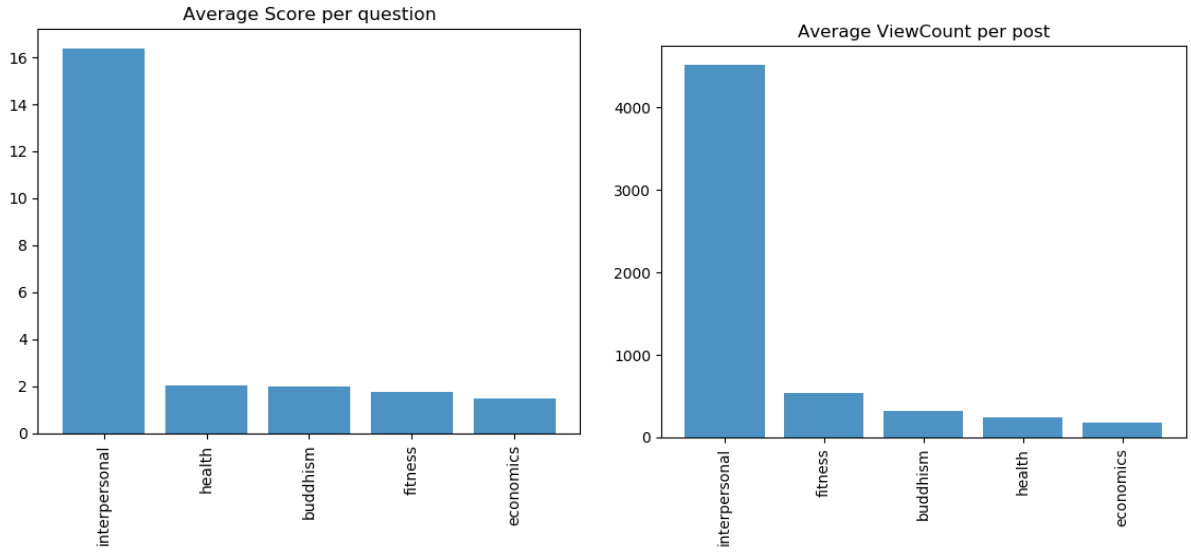
From the fora descriptives displayed in digure **??** displays violin plots for each forum. In it we see that `Score` across all fora except for Interpersonal is centred around

Figure 1: **Fora Descriptives**



Source: Own calculations in PySpark.

Figure 2: **Fora Descriptives**



Source: Own calculations in PySpark.

Interpersonal has a substantially higher average `Score` per question at over 16 compared to others which are just below, indicating that this community hands out votes the easiest. However, Interpersonal also has the highest `ViewCount`, signalling that it has the most viewing activity from both community-members and non-community members alike. When considering the composite variable, `Score/ViewCount`, Interpersonal interestingly comes last, showing that although they are overall the most generous at handing out up-votes, when compensating for how much viewing traffic the questions get the community actually has on average the least votes being cast out of the amount of views questions get (due to many non-community member views, many up-votes being offset by down-votes i.e. contentious issues??).

Needless to say, these communities appear to operate quite distinctly, making predicting community engagement by a single topic-based-model quite challenging, which would be contrary to the claim made by Ravi *et al.* (2014). The above average figures only display one dimension however, so we further graph the density curves for all three variables in figures blah through blah.

3.3 A Clear Definition of the Response Variable

3.3.1 The `Score` Variable

In order to robustly define a response variable capturing community engagement, there are certain aspects of the data and functioning of the StackExchange sites that should be discussed.

First of all, although questions on all StackExchange sites being open to the public, posting a question in a community requires registration with an email address and a username. Once registered, users start with a *reputation* level of 1 (https://meta.stackexchange.com/questions/7237/how-does-reputation-work). The reputation levels key to my analysis are laid out as follows:

- 15: Users are allowed to "up-vote" questions and answers
- 125: Users can "down-vote" questions and answers
- 1000: Users can edit any question or answer.

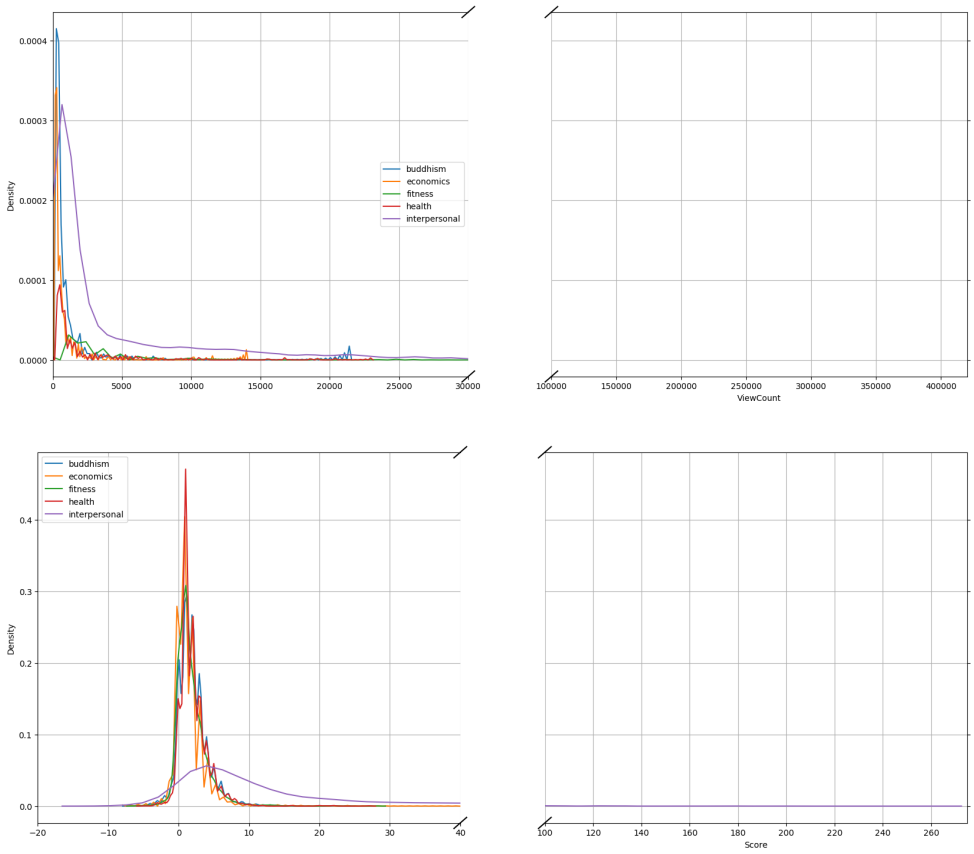
One factor is that there seems to be a less-than-full consensus of when exactly to up- or down-vote[1] despite general guidelines on StackExchange sites stating that up-votes should be given if a question shows prior research, is clear and useful, and down-voting the opposite.

One possible confounding factor for the response variable that is worth considering is that questions can be edited, not only by the original poster, but by anyone with a level of reputation of 1000 or more. General cross-community guidelines for editing include addressing grammar and spelling issues, clarifying concepts, correcting minor mistakes, and adding related resources and links. The concern here is that users could vote, comment and answer on substantially different questions over time as a question is edited from it's original form. **The simplifying assumption that I make here is that most edits, if any at all, would happen quickly as moderators are made aware of offending questions and thus the majority of views and votes would happen on final, edited questions. I therefore choose final edited question content to predict on.**

The contrasting reputation levels for up- and down-voting privileges (15 and 125 respectively) also lead to a `Score` variable that is highly negatively skewed, making it more likely that questions will have a higher `Score` and giving the appearance that most questions are highly valued by a community. The density curves of the `Score` variable per community in figure 3 below confirm this. The negative skewness of all 5 curves is clearly evident, and all except Interpersonal are steeply centred just after 0.

[1] https://meta.stackexchange.com/questions/12772/should-i-upvote-bad-questions

Figure 3: **Density Plots**



Source: Own calculations in PySpark.

Despite the negative skewness, because I am predicting a continuous response variable as opposed to binary, **I believe this won't be an issue.** It is just the way that the site functions and thus my prediction model will not not be altered to try and force the data to my whim.

3.3.2 The `ViewCount` variable

A really interesting aspect of the data and of the functioning of the StackExchange sites concerns the `Score` and `ViewCount` variables. **As mentioned**, only members that have registered with the community are able to up-vote and down-vote, thus contributing to the `Score`, but owing to all questions being open to the public, `ViewCount` variable registers views from 1) registered users that can vote, 2) registered users that can't vote due to a reputation level below 15), and 3) non-registered members.

This caveat influences the methodology of Ravi *et al.* (2014) in two ways. First, they decide to use a composite response variable, `Score/ViewCount`, since thise would move in the direction of taking the proportion of users that have decided to vote out of the total number into account (i.e. the percentage of voters that decided to up-vote a question could be small in comparison to the total amount of users that viewed the question). As they rightly point out, considering `Score` alone might lead to conflating popularity with their goal of measuring question quality, since a higher `ViewCount` would definitely bias the `Score` variable upwards owing to the unsymmetrical reputation privileges.

That said, here we have a look at the correlations between the `Score` and `ViewCount` variable:

Table 2: **Score and ViewCount Correlations Across Fora**

Forum	**Score and ViewCount Correlation**
Buddhism	0.41
Economics	0.67
Fitness	0.26
Health	0.21
Interpersonal	0.87

Source: Own calculations in PySpark.

I like to place this challenge in a framework of "within-community" engagement, and "outer-community engagement". Votes from community members, and consequently the `Score` variable, is purely a within-community metric since you have to be registered with the community to contribute to this variable. `ViewCount` on the other hand, is both a within- and outer- community engagement variable, since it does not distinguish voting or non-voting status when registering question views. I choose not to mix these two metrics and focus solely on within-community engagement through the `Score` variable - although this may mean popularity may be entangled in this response variable I have chosen, I would assert that popularity is in itself a measurement of community engagement and users would really only be interested in this final `Score` prediction rather than something like `Score/ViewCount`.

The second methodological adjustment that Ravi *et al.* (2014) make with their data is to only consider questions above a certain minimum `ViewCount` threshold. Their reasoning behind this is so that they can be more confident of the final dataset containing questions that have been viewed by qualifying users that can vote, or in other words their claim is that questions with higher `ViewCounts` have a higher probability of having been seen by community members able to vote.

I believe this is a false claim, since one could just as easily argue that new questions that begin with a low `ViewCount` are more likely to see engagement from proactive community members, especially if these questions doesn't generate enough webpage activity to rise as the top hit for search engines (which would lead to more non-community member activity contribution to views). Since there is additionally no data on the distribution of qualifying and non-qualifying user contributions to the `ViewCount` variable, therefore I opt to not disregard any questions below a certain `ViewCount` threshold.

3.3.3 Final Response Variable

Table 4 displays the titles of a selection of community questions with the highest and lowest `Scores`, i.e. a selection of the "best" and "worst" questions according to the methodology I have chosen.

Table 3: **Highest and Lowest Scored Questions Across Fora**

Forum	Score	ViewCount	Title
Buddhism	24	7228	Is low self-esteem a Western pheno
Buddhism	-7	103	Who remembers the Buddha?
Buddhism	-7	447	Why are buddhists hostile?
Economics	61	14055	What are some results in Economi common sense?
Economics	-7	179	What is feminist economics?
Fitness 28	12376	Why does one person have lots of stamina and another doesn't?	
Fitness	-6	54	Gaining fat for muscles-stomach fa
Health	27	4364	What are known health effects of s
Health	-5	35	Do "whole body jolts" experienced licking one's hear, chalk screeching
Interpersonal	265	32147	What to do if you are accidentally
Interpersonal	-9	1327	How can I tell if family members c
Interpersonal	-9	937	How to tell employees that I don't

Source: Own calculations in PySpark.

Table 4 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. We can see that this is now validated.

Table 4: **Descriptives for the Score Variable**

Forum	Mean	Standard Deviation
Buddhism	24	7228
Economics	24	7228
Fitness	24	7228
Health	24	7228
Interpersonal	24	7228

Source: Own calculations in PySpark.

3.4 Model

3.4.1 Train/Test Split

Let q_i denote question i out of all questions Q for a given forum. I split the datasets into a training set Q_{train} (50%) and a testing set Q_{test} (50%), each with 1 560 questions. **I choose a 50/50 train/test split because I believe that the size of the datasets allows for enough training data.** The standard deviations of a random splitting of training and testing sets is displayed in table 5 below.

Use SD or σ?

Table 5: **Random Train/Test Split Standard Deviations**

Forum	Train SD	Test SD	% Difference
Buddhism	2.1	2.27	8.1
Economics	1.86	3.34	79.57
Fitness	2.18	2.1	-3.67
Health	2.11	2.01	-4.74
Interpersonal	22.37	24.96	11.58

Source: Own calculations in PySpark

We see in table 5 that the standard deviations of **one forum** are clearly distinct from the others.

However, when the train/test split is done so that the training set questions chronologically precede the testing set questions, there are substantial differences in standard deviations, as shown in table 6.

Table 6: **Temporal Train/Test Split Standard Deviations**

Forum	**Train SD**	**Test SD**	% Difference
Buddhism	2.42	1.82	-24.79
Economics	3.16	2.1	-33.54
Fitness	2.17	2.09	-3.69
Health	1.96	2.16	10.2
Interpersonal	25.97	20.47	-21.18

Source: Own calculations in PySpark

This shows that the data is heterogenous with regard to time, either due to how the communities have evolved over time or how questions have evolved.

I use the random train/test split for the first part of the analysis and the temporal split for the second. This touches on a point that was not considered in Ravi *et al.* (2014) nor in previous research to my knowledge - the temporal nature of online Q&A questions. I believe that predicting `Scores` of future questions may prove a substantially more difficult task than just randomising the training and testing question sets.

3.4.2 Elastic-net Regularised Regression Model

I use elastic-net regularised regression to predict the score, denoted s_i, of each question using only features derived from the raw textual `Body` and `Title` independent variables, which I shall denote $\boldsymbol{x}_i'$. The learning objective can therefore be summarised as finding a coefficient vector $\boldsymbol{\beta}$ which minimises the Root Mean Squared Error:

$$\underset{\boldsymbol{\beta}}{\text{minimise}} \quad \sqrt{\frac{1}{|Q_{\text{train}}|} \sum_{q_i \in Q_{\text{train}}} (s_i - \boldsymbol{\beta}\boldsymbol{x}_i')^2 + \Psi} \tag{3.1}$$

where

$$\Psi = \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \tag{3.2}$$

is the elastic net penalty term. **WHY CHOSE RMSE AS METRIC**

In this term, λ is the regularisation parameter, α is a weighting coefficient for the L_1 and L_2 norms of the input variables, corresponding to the lasso and ridge penalties respectively. **MORE**

LASSO better when there are variables that are useless (they get shrunken to 0), RIDGE better when all are useful because it will shrink parameters but not eliminate.

I use 2-fold cross validation - 2 because increasing the number of folds did not lead to large gains in RMSE reduction over models in general, and also drastically increased computation time.

3.4.3 Question Content

A number of preprocessing steps are applied to the `Body` and `Title` to obtain the final features $\boldsymbol{x}_i'$ that are discussed subsequently - I parsed the HTML of the question content in the `Body` variable, tokenised (with punctuation) both the `Body` and `Title` texts, removed English stopwords and stemmed tokens using Porter-stemming (**???**).

I first extract features relating to the length of questions' `Body` and `Title`, i.e. token count, sentence count and character count. Then, the actual unigram text of the question `Body` and `Title` are used as features in the form of term frequency – inverse document frequencies (TF-IDF). **MORE Since Ravi *et al.* (2014) do not use higher order ngrams, I also stick to unigrams, resulting in quick and compact learning.**

3.4.4 Topic Modelling

I train an LDA model globally over all questions in Q. I use the online LDA learning framework in the Pyspark `pyspark.sql.ml` package to generate topic distributions over words for each question and add these as model features. This results in features made up of weights θ_{qt} for a topic t in a question q, and $\theta_{qt} = P(t|q)$.

I choose $K = 10$ topics

Online LDA works like this

3.5 Temporality

Since a large concern of my analysis (that hasn't been considered before) is temporality of the data, I explore a temporal splitting of I have two sets of results - non-temporal results that mirror Ravi *et al.* (2014), and temporal results where we try and predict future question `Scores`.

I split that set of questions, Q, into a training set Q_{train} and testing set Q_{train} using a point in time to mimic the reality of employing this tool at a certain point in time with historical training data. With the chosen date of **1/1/2017**, the ratio of Q_{train} to Q_{train} are ..... for ..... respectively.

**This addresses something that has not been considered in the literature and introduces a temporal element into the analysis.

We look at the descriptive statistics of the train/test split:

Note however that I have not included a temporal element to my model, so if there are some time-series trends in the data (to do with the struture of the websites changing etc.), then the temporal prediction will be poor.

4 Results

4.1 Random Train/Test Split

To establish a baseline for the predictive performance of the models, table 7 displays training and testing RMSE values across fora for a model that predicts the constant mean of the training set for every question in the testing set.

Table 7: **Constant Mean Model**

Forum	**Train RMSE**	**Test RMSE**	**Time (s)**
Buddhism	2.10	2.27	0.50
Economics	1.86	3.34	0.58
Fitness	2.18	2.10	0.52
Health	2.11	2.01	0.49
Interpersonal	22.36	24.96	0.58

Source: Own calculations in PySpark

The values in the Test RMSE column in table 7 are considered the low benchmarks that future models must improve upon. Interestingly, test RMSE is lower for all communities than train RMSE, thus it appears that there is substantially more noise in the training sets (remember that the training sets are older questions as well).

The RMSE is different per community owing to the different standard deviations of the data as a whole seen in the EDA...

Table 8: **ViewCount Model**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)
Buddhism	1.95	2.03	10.57	7.07
Economics	1.70	2.55	23.65	4.36
Fitness	2.06	2.13	-1.43	5.70
Health	2.06	1.98	1.49	5.61
Interpersonal	10.98	12.72	49.04	5.94

Source: Own calculations in PySpark

Table 8 is the high benchmark owing to the strong correlations seen in table 2. It is also vacuous, since the final `ViewCount` of a question is not available for new questions.

Table 9: **Length Model**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)	Elastic Param	Reg'tion Param
Buddhism	2.08	2.25	0.88	14.04	0.01	0.01
Economics	1.86	3.33	0.30	10.34	0.01	1.00
Fitness	2.17	2.09	0.48	7.60	0.01	1.00
Health	2.09	2.01	-0.00	7.93	1.00	0.01
Interpersonal	22.31	24.88	0.32	14.51	1.00	1.00

Source: Own calculations in PySpark

Table 9 shows mild gains from approximately 0.3% above the constant mean benchmark in the Test Gain column. We also see that the Interpersonal forum differs from the others in that the grid search found a regularisation parameter of 1 to be the most optimal, implying ...

Table 10: **Unigram Textual Model**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)	Elastic Param	Reg'tion Param
Buddhism	2.10	2.27	-0.00	221.30	1.0	1.0
Economics	1.86	3.34	-0.00	201.97	1.0	1.0
Fitness	2.18	2.10	-0.00	190.30	1.0	1.0
Health	2.11	2.01	-0.00	186.69	1.0	1.0
Interpersonal	15.06	25.87	-3.65	358.41	1.0	1.0

Source: Own calculations in PySpark

The results of using unigram text of question titles and bodies is displayed in table 10. This

model struggles particularly, only predicting means of the training set questions' `Score` for every community except for Interpersonal, where it actually performs worse than just predicting the mean by 3.5%. **What does this imply???**

Interestingly, the grid search gives an elastic parameter of 1 and regularisation parameter of 1 for all models.

Table 11: **Global and Local Topic Model**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)	Elastic Param	Reg'tion Param
Buddhism	2.05	2.25	0.88	11.36	0.01	1.0
Economics	1.86	3.34	-0.00	9.98	1.00	1.0
Fitness	2.18	2.10	-0.00	12.33	1.00	1.0
Health	2.11	2.01	-0.00	14.49	1.00	1.0
Interpersonal	22.31	24.95	0.04	11.47	1.00	1.0

Source: Own calculations in PySpark

It looks like predicting the score variable using just the textual content of questions is not going well. What we have garnered is that the fora are very heterogenous, having seen large differences in both the descriptive statistics and predictive results - different parameters come out as optimal for the length and topic models.

Table 12: **Length and Topic Model**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)	Elastic Param	Reg'tion Param
Buddhism	2.04	2.24	1.32	11.24	0.01	1.0
Economics	1.86	3.34	-0.00	8.96	1.00	1.0
Fitness	2.18	2.10	-0.00	8.26	1.00	1.0
Health	2.08	2.00	0.50	8.27	0.01	1.0
Interpersonal	22.26	24.87	0.36	8.44	1.00	1.0

Source: Own calculations in PySpark

4.2 Temporal Results

Table 13: **Length and Topic Model For Temporal Train/Test Split**

Forum	Train RMSE	Test RMSE	Test Gain (%)	Time (s)	Elastic Param	Reg'tion Param

Buddhism	2.35	2.00	0.99	17.09	0.01	1.00
Economics	3.08	2.25	-1.35	11.54	0.01	1.00
Fitness	2.11	2.13	-0.47	9.23	1.00	0.01
Health	1.93	2.16	-0.00	8.43	0.01	1.00
Interpersonal	25.57	22.02	-0.55	8.27	1.00	1.00

Source: Own calculations in PySpark

While I do not employ time-series models, I leave it to further research to incorporate a way to also "remember" which questions are good, so that in future there are no duplicates.

EVEN AFTER GETTING RID OF A SUBSTANTIAL AMOUNT OF DATA FOR CERTAIN DATASETS BY ONLY USING THE LAST TWO YEARS WORTH OF DATA, THE TEMPORAL MODEL STILL STRUGGLES SUBSTANTIALLY.

5 Limitations

Different motivations behind voting.

Different interventions from StackExchange sites (nudges introduced already.)

One aspect of this research that stands out as an area for further research is the fact that only one target variable was considered (i.e. `Score`/`ViewCount`) as a measurement of community interaction, whereas in reality there are others already available in the data. There are metrics recording how many interactions a question receives, such as `AnswerCount` (the number of answers for a question) and `CommentCount` (the number of comments for a question), which all signify at least some engagement with a question, although whether this is positive or negative engagement is unknown. In response to this, one could construct a variable relating to the linguistic sentiment of the answers (not comments, since comments need not be directed at the original questioner), however the subtleties of identifying sarcastic and condescending answers and comments might be overly difficult, especially since communities would value pleasant critical feedback.

Another variable that is a direct indication of questioner satisfaction is whether they deem an answer to have successfully addressed their question, which is recording in the variable `AcceptedAnswer`. This variable is not without its own issues, since users may find utility from multiple answers and neglect to formally select an accepted answer at all, biasing the number of formally solved questions downwards and confounding the response variable. Furthermore, answers are commonly posted as comments and vice-versa (see https://meta.stackexchange.com/questions/17447/answer-or-comment-whats-the-etiquette), and this too would confound the predictive results for this variable. Comments being posted as answers (i.e. "clogging up" the list

of answers), can be a case of users who don't have the required level of reputation to comment yet or a case of users chasing reputation points by using jokes, which obscurs the reputation measurement as users get voted up for being humourous rather than their expertise. Treating this variable as the target variable also situates the research problem in terms of exclusive utility to the user, whereas the `Score` variable is a more broader measurement of how the community values questions, which in turn should translate into utility for the questioner. One assumption that would mitigate issues surrounding the `AcceptedAnswer` variable would state that the discussed anomalies are not common enough and are not biased to specific posts with an even and randomly distribution over the data it would not significantly effect the results.

One last response variable for consideration is the number of times a post is edited, the `EditCount`. This variable could have two implications however - more edits signify more effort needed to bring the question in the desired state (i.e. it is inversely proportional to positive community engagement), or more edits signify more energy willingly devoted to improving the question because it will add value to the community (and thus it is directly proportional to positive community engagement).

Temporal aspect

6 Recommendations for Further Research

As has been discussed, there are other response variables for consideration, each with their own merits and disadvantages, however further research could investigate these and address the issues surrounding each response.

As mentioned, one limitation is that questions can be edited not only by the original poster, but also by anyone with 2000 reputation or more. One suggestion for further research would be investigating average times-taken for events such as edits, answers, votes and views to ascertain if the assumption of most votes occurring before edits is permissable (NO DATA ON THIS THOUGH).

Another is that the above model does not take into account the temporal nature of questions - i.e. a good question that is asked will be received positively by the community, however if a very similar question is asked later on, the community will see that as "lack of prior research" and will respond negatively. This analysis is but just a first step in accurately predicting positive community reaction, so further research could address the temporal model.

7 Concluding Remarks

The aim of this research was to predict the range of positive/negative community engagement that questions elicit. I believe that no prior research has endeavoured with the methodology here in this respective framework to predict and capture community engagement. At the very least, the research here has improved upon the extent of how community engagement can be ascertained from online Q&A communities, and has yielded insight into how homogeneously community engagement exists over diverse communities with various subject matter. I believe that using this tool, online Q&A users will be assisted in improving their submitted questions which will enhance the productivity of all online Q&A communities wholly. Furthermore, room exists for implementation on any assortment of Q&A sites, counting Massive Open Online Courses.

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.

Allamanis, M. and Sutton, C. (2013) 'Why, when, and what: Analyzing stack overflow questions by topic, type, and code', in *2013 10th working conference on mining software repositories (msr)*. IEEE, pp. 53–56. doi: 10.1109/MSR.2013.6624004.

Anderson, A. *et al.* (2012) 'Discovering value from community activity on focused question answering sites: a case study of stack overflow', in *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 850–858. Available at: http://dl.acm.org/citation.cfm?id=2339665.

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.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, 3, pp. 993–1022.

Chiang, D. *et al.* (2010) 'Bayesian Inference for Finite-State Transducers', in *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. Association for Computational Linguistics (June), pp. 447–455. Available at: http://www.isi.edu/{~}sravi/pubs/naacl2010{_}bayes-fst.pdf.

Fligner, M. and Verducci, J. S. (1986) 'Distance based ranking models', *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3), pp. 359–369.

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.