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 crucial to how we examine, learn from and critically analyse the world around us. The evolution of the world wide web and the technologies that have emerged have given us an unprecendented ability to engage with and learn from individuals around 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. The domain of online question-answer (Q&A) communities is one setting where relevant and well-researched questions is of particular importance, since domain expert resources are generally scarce compared to the volume of new questions (a problem known as information overload). This research builds on the small amount of work aimed at questions in online Q&A communities, and begins to address the problem of information overload by analysing a diverse range of questions from the StackExchange family of Q&A communities. Using only textual question content available at the time questions are initially submitted, I construct models to predict on the community-granted score for each question as a measurement of community engagement. I find little to **no improvement in prediction metrics after employing feature engineering techniques sourced from the relevant literature across all fora.** Having taken the step to predict a continuous proxy measurement of online community engagement and begun **addressing temporal issues**, I belive my **unique and original** research shows that there is still much to be done to predict online community engagement objectively and effectively, especially when considering the **chronological nature of online Q&A data**. 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. (317)

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. One avenue of online interaction that has become an extremely popular way in which users share knowledge about diverse and nuanced subject matter is question-and-answer (Q&A) websites such as Yahoo! Answers, Quora, the StackExchange family and forums of Massive Online Open Courses (MOOCs). These websites serve as dynamic, engaging platforms where users seek answers to and discussions on complex and technical questions that modern search engines are evidently yet unable to fully address.

In these online Q&A fora, producing relevant, well-researched and high-quality questions is especially valuable not least since these platforms suffer in particular from a low ratio of expert resources to volume of new questions - a problem known as *information overload* (**???**). The overarching hypothesis of this research is that if questioners in online Q&A fora were provided with information specifically related to how well their questions will be received by communities, then they could iterate to "increase the signal" of their questions before exerting demand on community resources, thereby mitigating the problem of information overload. This would no doubt benefit questioners as they become better able to garner expert answers to their improved questions, but also benefit entire communities as overall functioning and efficiency is improved community-wide.

Addressing information overload in online Q&A is a non-trivial problem however, since providing predictions of *community engagement* to questioners in real time requires that only the information available when new questions are formulated can be used as features, i.e. question content as opposed to other features such user characteristics, final webpage viewing statistics etc. Furthermore, final predictions given to questioners would also ideally be highly granular and direct questioners towards how best to improve their questions (leading to an intersection with the vast literature on recommendation systems), however the undeniable first step, and what this research aims to achieve, would be to ascertain if community engagement in online Q&A fora can actually be predicted with some measure of accuracy.

The broad research question for this paper can therefore be summarised as the following:

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

While there is a substantial amount of literature that has addressed Q&A communities, interestingly the focus has been on identifying expert users and high quality answers rather than looked at questions, despite questions being the entry point for every interaction in communities. To answer the research question above I analyse a diverse range of Q&A fora from the StackExchange family of communities, and draw heavily on prior research done by Ravi *et al.* (2014) on question quality in online Q&A fora.

Using only the textual content of questions, I predict on each question's community-assigned `Score` - an aggregation of all community *up-votes* and *down-votes* which I accept as an objective and comprehensive metric for community engagement. In line with the analysis of Ravi *et al.* (2014), I employ Latent Dirichlet Allocation (Blei, Ng and Jordan, 2003) to engineer latent topic features from question content for predictions. I use elastic-net regularised regression for the learning task and evaluate models using root-mean-square error (RMSE).

It should be highlighted that the goal of this research is quantitative prediction rather than qualitative, causal or inferential analysis. While I will briefly touch on the differences between predictive model results across the communities I analyse, I leave it to further research to address more precisely the *how* and *why* of community engagement in online Q&A communities, rather than just the *if* that is explored here. To my knowledge this research is the first of its kind to test latent topic models on a continuous and objective measurement of online community engagement, analyse a diverse range of communities, as well as have a real-life use-case, resulting in a practical and unique contribution.

My findings show that there is still much work to be done to accurately predict community engagement in online Q&A fora. I find that models that include features derived from the lengths and topics of questions do **NOT** perform better than a baseline of just average `Score` prediction from the training question set. I also find that models across fora have **varying** levels of performance, **providing evidence** against claims in Ravi *et al.* (2014) that topic models are applicable to different online Q&A settings.

Lastly, as a progressive step forward in the research, I evaluate the best performing model using a temporal train/test split, taking into account the chronological nature of online Q&A questions. Here I find **almost no gain in predictive performance**, leading me to believe that in order to accurately predict future community engagement, models need to incorporate temporal and time-series elements.

In the following section I discuss relevant literature in more detail. This is followed by section 3 which discusses the data, explores and validates my choice of `Score` as an objective measurement of community engagement and describes the predictive model used. Section 4 presents and discusses the results, section 5 makes some recommendations for areas of further research and finally section 6 makes some concluding remarks.

2 Literature Review

2.1 Question-Answer Communities

There is a substantial collection of research that has investigated online Q&A communities. Prior work has addressed 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 in the form of a recommendation system (Wu, Wang and Cheng, 2008; Qu *et al.*, 2009; Szpektor, Maarek and Pelleg, 2013).

This research differs from this previous work on Q&A fora in two respects. Firstly, I focus on questions rather than user or answer characteristics, not only because they have received substantially less attention in the literature, but because it has been shown that question quality can substantially impact the quality of answers (Agichtein *et al.*, 2008). Questions in online Q&A fora are also the initial event that all community engagement follows from and thus maximising positive community engagement with questions will almost certainly improve the evolution and functioning of communities.

The second distinction from prior research is the framework in which this research is placed. I choose a framework of community engagement and interaction with user actions rather than the systems-based optimisation of question-answer routing and matching and instead concentrate on how questioners can be nudged to improve the content of their questions before encumbering community resources.

Community engagement is a rather broad term in literature ranging across fields and disciplines, but I have not found any literature relating to community engagement in the context of online Q&A fora.

With the promise of this real-life application which significantly benefit both questioners and communtities, it remains to be seen if community engagement in online Q&A fora can be successfully predicted. Owing to a large overlap between this goal and the literature on predicting

question quality in online Q&A fora, 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 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 is 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". I believe that community behavior is too diverse to be universally predicted by a single model, **thus this will be interesting to test in the results.**

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.

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

Table 1: **Details of Datasets**

Forum	Questions	Answers	Users	Site Age	Description
Stack Overflow	18m	28m	11m	11yrs	Q&A for professional and enthusiast programmers
Super User	420k	605k	795k	10yrs	Q&A for computer enthusiasts and power users
Math	1.1m	1.6m	567k	9yrs	Q&A for people studying math at any level
Cross Validated (Stats)	143k	143k	209k	9yrs	Q&A for people interested in statistics
English	107k	250k	267k	9yrs	Q&A for English language enthusiasts

Owing to the size of the datasets, I process and analyse the data with PySpark, a Python API for the open-source cluster-computing framework Apache Spark. In the interests of transparency and reproducibility, the entire PySpark codebase for the processing and modelling of the data done can be found at https://github.com/BCallumCarr/msc-lse-thesis/.

The data from the five selected fora is downloaded, decompressed and converted to Parquet format. The following variables are of interest to my analysis from the data:

- `Score`: The difference between registered-user granted up-votes and down-votes for a question
- `ViewCount`: A counter for the number of page views a question receives (form both registered and non-registered users)
- `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
- `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)
- `AcceptedAnswerId`: Indicates which answer the question-asker selects as accepted (questions only)
- `ClosedDate`: A date variable indicating if a question was closed (questions only)

There are two options for selecting the final data I wish to analyse: Selecting an equal number of questions from communities with the datasets spanning different lengths of time, or selecting a common length of time but then having datasets with different sizes. This is our first insight into how the temporal nature of online Q&A data can complicate analyses. Since the main goal of this research is not to compare communities (although a natural comparison will evidently surface), I choose to mitigate as much temporal nature in the datasets by choosing a relatively short and uniform time period in which to extract the final data. The combination of analysing data over a uniform time period, and also trimming away recent questions that would not have existed in communities long enough to garner sufficient votes and views results in a final dataset that should have little to no temporal nature and bias.

The dates I choose to examine the datasets are from the **1st of September 2010 to the 1st of September 2011**, since this is from the start date of Maths, Stats and English. This trims the initial number of questions down to **1 113 802** questions, or more specifically 1 042 477 for StackOverflow, 40 589 for SuperUser, 18 131 for Math, 8 537 for English and 4 068 for Stats.

It should be noted that the variation in the length of time periods from which questions were extracted across fora may complicate comparison between fora if the data exhibit temporal effects and trends - this will be explored in more detail a bit later. I now move onto formalising the measurement of community engagement that will be predicted on.

3.2 A Measurement of Community Engagement

3.2.1 An Exploration of Community Engagement Variables

There are a number of ways that online Q&A community members interact...

One aspect of this research that stands out as an area for further research is the fact that only one target variable, the question `Score`, was considered as a measurement of community engagement, 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).

**I won't use the `AcceptedAnswer` variable since I believe it unreliable, and providing less detail as a binary variable than continuous variables like the `Score` and `ViewCount`.

3.2.2 An Exploration of Community Engagement Variables

The `Score` and `ViewCount` variables are the best candidates for community engagement and thus we perform exploratory analysis on them:

Table 2: **Descriptives for the ViewCount Variable**

Forum	Count	Mean	SD	Min	Max
Buddhism	3120	316.16	737.08	10	21498
Economics	3120	178.24	646.26	2	14055
Fitness	3120	534.35	2426.63	8	78829
Health	3120	240.40	1128.95	2	23098
Interpersonal	3120	4520.74	7492.72	6	79049

Source: Own calculations in PySpark.

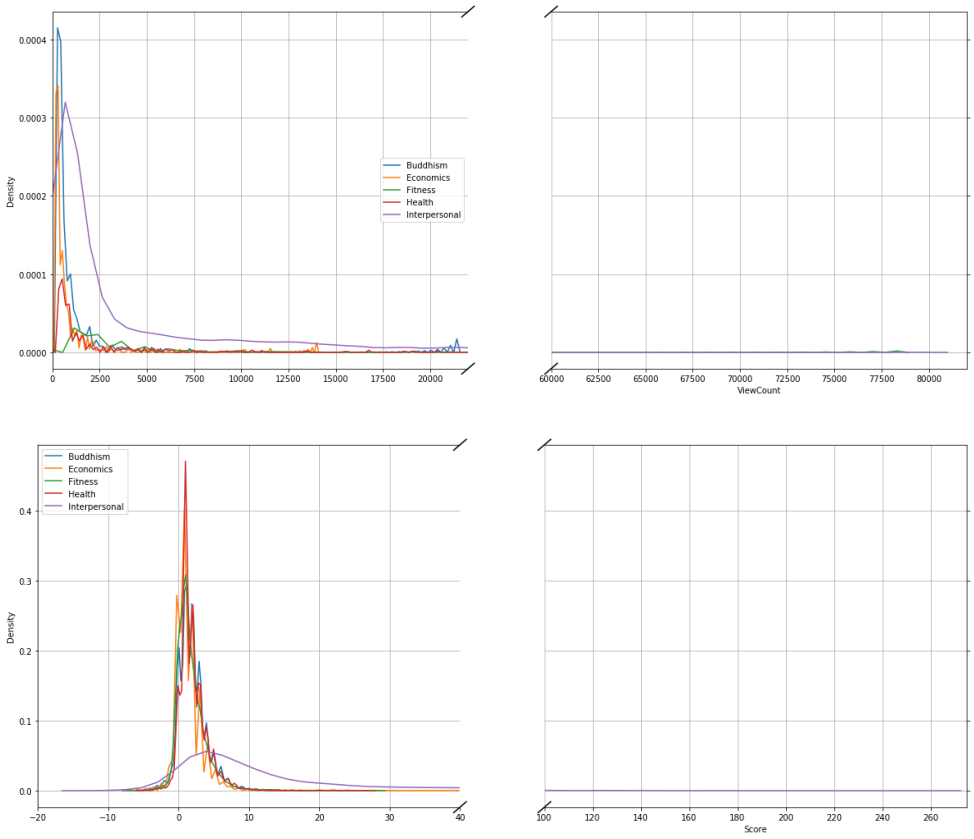
Table 3: **Descriptives for the Score Variable**

Forum	Count	Mean	SD	Min	Max
Buddhism	3120	2.01	2.19	-7	24
Economics	3120	1.47	2.70	-7	61
Fitness	3120	1.77	2.14	-6	28
Health	3120	2.05	2.06	-5	27
Interpersonal	3120	16.38	23.70	-9	265

Source: Own calculations in PySpark.

From the above descriptive tables we seee that there is much heterogeneity across the `ViewCount` and `Score` variables. Most notably, the averages and variances of `ViewCount` and `Score` from Interpersonal are orders of magnitude higher than other fora. This shows begins to shed light on how distinctly these communities appear to operate quite distinctly, either due to community behaviour or questioner behaviour - not community size because as we see in table 1, this appears to not be related to the **descriptive statistics above. As feels trivially true, if communities value certain aspects of questions differently or behave differently, predicting community engagement with a universal model feels quite the challenging.** For further information on these variables, we plot density plots of the `Score` and `ViewCount` variables:

Figure 1: **Density Plots**



Source: Own calculations in PySpark.

In both density plots we see much visual confirmation of the differences between fora, most notably the **outlier-ish data points for both the `Score` and `ViewCount` variables from the Interpersonal forum.** The distributions of both variables across all fora also appear to be significantly uneven and negatively skewed. This seems to coincide with Benford's, Zipf's and Pareto's laws (https://workplace.meta.stackexchange.com/questions/5018/massive-viewcount-difference).

There is also the matter of the websites functionings themselves. Although questions on all StackExchange sites are open to the public, posting a question in a community requires registration with an email address and a username and once registered, users start with a *reputation* level of 1 (https://meta.stackexchange.com/questions/7237/how-does-reputation-work). Key reputation levels include:

- 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.

At least for the `Score` variable, the contrasting reputation levels for up- and down-voting privileges (15 and 125 respectively) would bias this variable to be negatively skewed, as seen in figure 1, making it more likely that questions will have a higher positive `Score` rather than a negative one.

Asides from the heterogeneity across fora, it appears that both variables exhibit similar distributions, and this leads us to discuss and identify what they represent, and how they could be used to measure community engagement.

This leaves us with the `Score` and `ViewCount` variables for analysis.

3.2.3 `Score` versus `ViewCount` for Measuring Community Engagement

With our two candidates for measuring community engagement, we now move onto a deeper discussion of how these variables arise. Only members that have registered with the community are able to up-vote and down-vote and thus contribute 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 their goal of predicting "question quality" heavily. They decide to use a composite response variable, `Score/ViewCount` stating that considering `Score` alone might lead to conflating popularity with question quality, since a higher `ViewCount` would definitely be linked to a higher `Score` variable upwards owing to the unsymmetrical reputation privileges. This is easily seen when looking at a table of correlations between the `Score` and `ViewCount` variables across fora in table 4 below:

Table 4: **Score and ViewCount Correlations Across Fora**

Forum	Correlation
Buddhism	0.41
Economics	0.67
Fitness	0.26
Health	0.21
Interpersonal	0.87

Source: Own calculations in PySpark.

In the table above we see correlations ranging from 0.21 for Health, to 0.87 for Interpersonal. Thus it appears that the `Score` and `ViewCount` are indeed inexorably linked, but **we can dive deeper into what they actually represent**. The framework I would like to develop is that of *within-community* engagement versus *outer-community engagement*. I assert that voting by community members, and consequently the `Score` variable, is purely a within-community metric because users are required to commit and register with a community to contribute to this variable. `ViewCount` on the other hand, can be defined as both a within- and outer- community engagement variable, since it does not distinguish voting or non-voting status when registering question views.

I decide to mix focus solely on within-community engagement in my methodology and analysis, and therefore use on the `Score` variable as the response to be predicted. While this may imply that what I am predicting is also **popularity** of questions, I believe that popularity is in itself

a measurement of community engagement. I further believe that in providing predictive information to questioners about their new questions, questioners would be more interested in seeing their final `Score` prediction rather than `ViewCount` or `Score/ViewCount`.

3.2.4 A Final Response Variable

Finally, table 5 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 5: **Highest and Lowest Scored Questions Across Fora**

Forum	Score	ViewCount	Title
Buddhism	24	7228	Is low self-esteem a Western phenomenon?
Buddhism	-7	103	Who remembers the Buddha?
Buddhism	-7	447	Why are buddhists hostile?
Economics	61	14055	What are some results in Economics that are both a consensus and far from 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 fat
Health	27	4364	What are known health effects of smoking e-cigarettes
Health	-5	35	Do "whole body jolts" experienced from things like tasting vinegar, a puppy licking one's hear, chalk screeching, etc. reach the median nerve?
Interpersonal	265	32147	What to do if you are accidentally following someone?
Interpersonal	-9	1327	How can I tell if family members consider my unvaccinated kids a threat?
Interpersonal	-9	937	How to tell employees that I don't mean my insults seriously?

Source: Own calculations in PySpark.

It appears that questions that are well-received by a community are genuine and discussion-promoting, whereas those received negatively by communities, have elements of sarcasm and insincerity in the sense that that are not actually looking for answers - i.e. `....` Social norms also appear to be prevalent in community reactions to questions - case in point being the unvaccinated kids question from Interpersonal.

3.2.5 Potential Methodological Issues

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.**

Another factor is community behaviour confusion - 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.

A 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.

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

3.3 Model

3.3.1 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 6 below.

Use SD or σ?

Table 6: **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 6 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 7.

Table 7: **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.

Note however that 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.

Since this analysis has already taken the first step from prior research to look more broadly at community engagement, and a continuous response in addition, potential temporality of the data will not be thoroughly explored in the form of implementation of time-series models, but will be touched on by comparing model results for a random train/test split versus chronological.

3.3.2 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.

3.3.3 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.**

3.3.4 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

4 Results

4.1 Random Train/Test Split

To establish a baseline for the predictive performance of the models, table 8 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 8: **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 8 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 9: **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 9 is the high benchmark owing to the strong correlations seen in table 4. It is also vacuous, since the final `ViewCount` of a question is not available for new questions.

Table 10: **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 10 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 11: **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 11. 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 12: **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 13: **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 Train/Test Split

Table 14: **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 USING ONLY A YEAR WORTH OF DATA in the early stages of the communitites' existence, THE TEMPORAL MODEL STILL STRUGGLES SUBSTANTIALLY.

5 Recommendations for Further Research

A number of areas for further research stand out from the methodology I developed here. Firstly, there are still many more complex features that can be derived from question content alone that were not included in the models in this research. Word-embeddings ((**???**)) might be **more** successful in predicting community engagement metrics.

As discussed **in detail** in section 3.2.2, there are also a myriad of other options for community engagement besides the `Score` variable, each with their own advantages and disadvantages. While I believe I thoroughly justified and validated my choice of the continuous `Score` variable as a comprehensive and objective response variable, not least because accurate predictions of it would be **highly useful information for questioners wishing to improve their questions**, a thorough exploration and predictive modeling of other measurements of community engagement with the models developing here would be extremely valuable.

Another area previously discussed that is ripe for further research is the editing of questions. As a reminder, 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 how much editing takes place over questions, in what average timeframe edits are completed compared to votes cast and views accumulated, as well as how evenly editing is distributed over questions. This research would then be able to test my assumption that most edits, if any, take place before the majority of votes and views are recorded.

There are also finer nuances regarding the functioning of the StackExchange sites, **some of which were** discussed in section 3.2.5. For registered users in various communities, there remains some confusion on when to up-vote and down-vote questions (https://meta.stackexchange.com/questions/12772/should-i-upvote-bad-questions). This links with how there are potentially vastly different motivations behind voting? Over time various communities have also implemented different interventions to nudge users to better formulate and structure their questions, i.e. reminders of doing prior research, including reproducible code for programming websites, and even going as far as to check that the `Title` of new questions do not match previous questions too closely for fear of allowing a duplicate question to be asked in the community. These nudges would no doubt affect the distribution and evolution of questions temporally in the data, and consequently affect metrics such as `Score` and `ViewCount`, which links with the next final recommendation for further research.

Most importantly as hinted throughout and demonstrated at the end of this analysis, temporality of the data is something that needs to be taken into account. This is at least true in the sense that questions which have existed longer in communities trivially would have had more time

to accumulate votes and views, but my further hypothesis is that the variation in community engagement metrics has decreased substantially over time as users and communities have refined how they permit and value certain questions - this new hypothesis of heterogeneity over time can be tested with variance equivalence testing for samples across time. All of this suggests that any model aiming to predict future community engagement in online Q&A fora must be expanded to include temporal effects and time-series elements.

Even if the temporality issue was solved however, another challenge is getting the model to pick up duplicate questions (which are ill-considered in all communities). This would mean instilling in the model that a question can be similar enough to previous questions for the model to learn that it's a good question, however not too similar so that the community perceives it as a duplicate, assumes a "lack of prior research" on the questioner's part and then reacts negatively to the question. As one can see therefore, there is still much work to be done in order to accurately predict future communtity engagement in online Q&A fora.

6 Concluding Remarks

The aim of this research was to predict the range of positive/negative community engagement that questions elicit, with the practical application of providing this information to questioners so that they can improve their questions before adding demand to a community. 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.

There is much heterogeneity in the data I have analysed, not only across fora but also over time. This, as well as the fact that I have attempted to predict on a continuous variable rather than binary, makes the problem of predicting community engagement from text-only data substantially more difficult, as can be seen from the poor predictive performance of the models employed. I leave it to future research to employ more sophisticated time-series models to capture temporal effects/features from the data.

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.