Automatically Rating Discussion Quality of Reddit Content

Jacob Zimmer

Abstract

Online discussion forums are social cyberspace channels that allow for the distribution of user-generated content and peer-to-peer discourse surrounding a shared, specified topic. While these forums contain large amounts of valuable information on a variety of topics, content ranking algorithms on these sites are often flawed, resulting in sub-optimal filtering of content that pushes low-quality content to the forefront. Current ranking systems place too much emphasis on 'easy-to-measure' metrics that do not reflect the true quality of a content submission.

1 Introduction

Online discussion forums are social cyberspaces that host channels allowing for distribution of user-generated content (UGC) and peer-to-peer discourse. Usually revolving around a specific topic or collection of related topics, these spaces represent important archives of knowledge surrounding numerous subjects. Such sights are often arranged in a threaded structure: The master forum contains many topic-focused sub-forums; sub-forums feature a collection of posts relating to the given topic; and posts may be home to any number of comments.

In recent years, these forum-style sites have exploded in popularity, becoming hubs of substantial cultural discourse for millions of people around the world - particularly Reddit, which Amazons Alexa.com currently ranks as the 5th most visited site in the US and 15th most visited worldwide. Reddits rapid rise in popularity can be attributed directly to its users, as it is unlikely that Reddit would have gained such notoriety without the extensive (and voluntary) work performed by individual users in the vein of community-building and content-generation. Thus, to say that the platform is only as good as the contributions of its users would not be off base. This dependence on the community extends even beyond content-generation, however, as the sites voting feature serves as the cornerstone of its user-centric design.

Given the importance of sustaining a strong user base to Reddits existing business model, finding ways to promote the best content is a key focus of the company. Several times in the past, Reddit has altered their scoring algorithm in order to do exactly that: however, the current ranking method continues to suffer from many of the same issues, resulting in a suboptimal experience for all users.

2 Problem Definition and Data

While Reddit has changed their scoring algorithm several times in the recent past in an attempt to improve the filtering of content on the site, their current scoring algorithm - the lower bound of the Wilson score confidence interval for a Bernoulli parameter - continue to suffer from the same flaws as the previous methods. As a result of focusing almost solely on using crude but easy-to-use metrics to rank submissions, such as peer votes and post time, the current system consistently promotes low-value content while undervaluing quality discussion. This problem is of unique interest Reddit's management team; As the expansion of online advertising on Reddit and similar sites has mirrored the growth of their user-bases, finding ways to improve the user experience of the platform has a direct monetary advantage. Implementing a smarter scoring system would push high-quality content to the forefront of site content, providing positive returns to users that promote platform loyalty via improved site experiences. This enhanced site experience could potentially drive significant user base growth and, in turn, an increase in demand for advertising on the platform. In this project, I seek to explore a novel

algorithm that considers the quality of a content submission's subject matter above all other scoring factors. In order to evaluate my algorithm, I build two logistic regression classifiers (one each for comments and posts) to evaluate the predictive power of each of my features.

The sample used in this analysis is comprised of a total of more than 200,000 comments and over 15,000 posts uploaded to Reddit during September 2018. This data is pulled from the Pushshift API through Google BigQuery. Due to the vast amount of activity that takes place on the Reddit platform, I have limited my scope to only comments and posts from three hand-selected subreddits. Choosing subreddits to focus on represented a unique distributional challenge; I felt it was important to ensure that the sample featured data from subreddits with varying attributes, namely: topic, level of moderation, ratio of text posts to link posts, and community norms (using meso community norms as defined in Table 2 and Section 6.2 of (Chandrasekharan et al., 2018)). Given these constraints, the three subreddits chosen are as follows: r/IAmA, where users can ask questions to be answered by a verified 'guest' (usually people with notable professional/academic histories or otherwise extraordinary life experiences); r/science, which describes itself as "a place to share and discuss new scientific research", and r/pokemon, which serves as a place for fans of the 'Pokemon' multimedia franchise to congregate and discuss the subject of their fandom.

This sample provides a good variation in frequent attributes across each class: r/IAmA is a general-interest subreddit that experiences a slightly above-average level of moderation, is comprised of mainly text comments and posts, and features norm violations centering around ad hominem attacks on other users and meme responses; r/science falls squarely between general-interest and niche, topically, and showcases heavy moderation and a high concentration of link posts. The norms of r/science focus on eliminating personal opinions & reactions, joke responses, and off-topic conversation. Lastly, r/pokemon is a niche-interest subreddit with minimal moderation, features a roughly even mix of text and link posts, and has norms focused on eliminating hedging language and criticism of other user's opinions. Those selections were then further sampled so that all comments in the sample were the children of posts in the sample, resulting in a sample containing: 5 posts and 30,397 comments from r/IAmA, 11 posts and 27,788 comments from r/science, and 46 posts and 14,004 comments from r/pokemon.

After pulling this data and sampling based on subreddit, I further limited the sample size to 398 unique comments and 371 unique posts, selected randomly from within the sample defined above. This sample size was chose such that it was small enough to reasonably be annotated by hand, but large enough to provide an adequate sample size for the classifier.

3 Related Work

There is a large amount of past work related to scoring algorithms for user-generated content, particularly related to online forums. In (Wanas et al., 2008), the authors selected 22 features to rate content quality on, grouped into 5 major categories: (1) Relevance, (2) originality, (3) forum-specific features, (4) surface features, and (5) posting component features (i.e. links). Using NLP methods to calculate these features, the researchers built a non-linear Support Vector Machine (SVM) classifier to classify content with a high, medium, or low seed rating. Achieving an accuracy of 50%, the authors found that these categories were sufficiently accurate to provide seed ratings for posts, and highlighted stopword removal, POS tagging, and phrase extraction as possible ways to improve accuracy.

The model I plan on implementing will follow a similar pipeline as that utilized in this research; the equations provided by the researchers for calculating feature metrics will be quite useful for my work. Yet, my model intends to score Reddit interactions based on the quality of the discussion it creates rather than the likelihood of heavy traffic. As such, my model will weight timeliness considerably less than the model implemented by (Wanas et al., 2008). Additionally, I will tweak certain aspects of their models, such as removing stopwords during relevance feature calculations, and include additional features, such as penalizing swear words in the surface features category.

In (Weimer et al., 2007), the researchers utilized a similar approach, splitting the features into 5 categories: surface features, lexical features, syntactic features, forum-specific features, and similarity features. While the surface, forum-specific, and similarity categories in this study seem to over-

lap quite well with the surface, forum-specific, and relevance categories used in (Wanas et al., 2008), (Weimer et al., 2007) also differ in their approach by selecting syntactic and lexical features. Taking into account things like swear words and POS tagging, the authors achieve an average accuracy score of 89.10%. Unfortunately, as this project has developed, I have realized that this research paper is not of significant value to my project. While the approach taken in the paper is similar that used in (Wanas et al., 2008) and to my planned model, their writing and model information lack the depth of information I need for this paper to truly be of use (since (Wanas et al., 2008) is more in detail and overlaps heavily). However, because of this similarity I will still consider this paper while deploying and tweaking my model. Additionally, I may adopt one of their suggested baseline scores as a baseline of my own in the evaluation phase.

Another related work discusses topic modeling in online forums, which I plan to implement in my algorithm to identify comment relevance. In his work, author Paul Ton uses an NLP pipeline that concludes by using Latent Dirichlet Allocation (LDA) to uncover latent topics within all posts. Splitting his analysis into thread-centric and user-centric approaches, he selects 40 and 25 topics, respectively. Ultimately, the study concludes that LDA may not be completely appropriate for subforums, since the resulting topics are difficult for humans to make sense of, yet can reveal promising topics using the thread-centric approach (Ton, 2017). In this article, the author uses topic modeling techniques, namely LDA, that I plan on implementing as a part of the Forum-Specific category. By using his approach to apply LDA to subforums, I can compare the topics discussed in a given post or comment to those seen across s sub-forum in order to inform my classifier more accurately about how relevant a submission/comment truly is to its hosting forum.

4 Methodology

In this analysis, I have implemented an approach founded on the concepts explored in each of these past related works. Similarly to the pipelines implemented in (Wanas et al., 2008) and (Weimer et al., 2007), I identified 3 high-level categories of post/comment features with the aim of capturing discussion quality in terms of both direct contribution quality and stimulation of further discussion. Containing a set of 9 features total, these categories are: (1) Relevance Features, (2) Surface Features, and (3) Discussion Features. The following sections will outline these categories in detail.

4.1 Relevance Features

One of the most important factors in measuring the perception of comment/post quality is the relevance of the submissions content to the topic of the subreddit and/or thread that it resides in. To measure this characteristic for both posts and comments, I implemented two separate approaches using concepts explored in (Wanas et al., 2008) and (Ton, 2017). In order to create a representation of each subreddits true subject-matter focus, I first created Bag-of-Words model representation for every comment, every post (post title and body only), every thread (post title, body, and all child comments), and every subreddit (using combination of all posts and comments in that subreddit). Using this representation, I calculate Relevance scores for posts and comments as follows:

4.1.1 Post Topic Score

After calculating these Bag-of-Words representations, I then perform Latent Dirichlet Allocation (LDA) on each of subreddit to identify the 10 topics that are most associated with that subreddit as a whole, as well as the 10 tokens that are most associated with that topic. Next, after performing LDA on each post thread Bag-of-Words and obtaining each threads 10 most associated topics and their descriptive tokens, I calculate the Topic Score of a post ps in subreddit s using the posts set of topic description tokens T_p, the subreddits set of description tokens T_s, and the Posts thread Bag-of-Words B_t as follows:

$$count(T_p \epsilon T_s) / \mid B_t \mid \tag{1}$$

4.1.2 Comment Subreddit Relevance Score

This score intends to capture the relevance of a comment to its containing subreddit in order to measure the comments relevance to the overall subreddit discourse. To compute the relevance of each comment to its subreddit, I first create a set of keywords for each subreddit by calculating tf-idf measures on the subreddit Bag-of-Words models and selecting the top 15 words from each.

The score for each comment is then calculated as follows, using a comment cs Bag-of-Words B and its respective subreddits keyword list K_r:

$$count(Tokens_B \epsilon K_r) / \mid B \mid \quad (2)$$

4.1.3 Comment Parent Relevance Score

This score intends to capture the relevance of a comment to the threads parent post in order to measure its relevance to the immediate conversation. This score is computed by comparing the overlap of the tokens in a comments Bag-of-Words and the Bag-of-Words representation B_p of its parent post:

$$count(Tokens_B \epsilon B_p) / \mid B \mid \quad (3)$$

4.2 Surface Features

Another determinant of discussion quality is the physical appearance of a submission. Poor formatting, offensive language, late posting, etc. are all examples of ways appearance affects quality: Readability is a key facet of discussion quality, and a comment cannot be effectively contributing to discourse if it is not or cannot be read. The features in this category are calculated as follows:

4.2.1 Length Score

The length of a post, while seemingly a basic metric, is quite important to how valuable a post or comment. Longer comments can, by their very nature, contain more information than shorter submissions. This score is calculated as follows: For posts, score is the total length in characters of a posts combined title and selftext, divided by the average length of all posts on the subreddit; For comments, this is the total length of its body divided by the average length of all comments in its containing thread (all children of its parent post).

4.2.2 Timeliness Score

In order for a comment or post to contribute to discussion, it must first be read by other users; If that comment or post is submitted long after a thread has gone silent, the information it contains is essentially irrelevant to subreddit discourse since there are no other users present to take interact with it. As such, this score accounts for the time that a post or comment was submitted in order to gauge the likelihood of other users interacting with it. This score is calculated as follows: For posts, I compute the average time of day that posts are submitted to the subreddit, and calculate the score as the absolute difference in hours of between a posts creation time and the average creation time for posts on that subreddit; For comments, I take a thread-level approach, calculating each threads average time difference between comment creation time and parent post creation time. Then, for each comment, I compute the difference between the comments creation time and the parent posts creation time, and dividing by the average thread comment creation time.

4.2.3 Capitals Score

Beyond size and timing, discussion contribution quality is also impacted by the direct physical appearance of a comment or post. For example, submissions that are presented in all capitals are often (1) less likely to contain valuable information, and (2) more likely to be ignored by later users. Thus, this score serves to account for such cases, and is calculated for both posts and comments as the count of capitalized characters in the body (or combined title and selftext for posts) divided by the number of words in the submission text.

4.2.4 Curses Score

In a similar vein as the capitals score, a post or comment containing curses and insults is less likely to contain valuable information and more likely to be ignored by other users - negatively impacting both the direct contribution quality and the probability of further discussion generation. For both comments and posts, this score is calculated as the count of curse words in the body (or combined title and selftext for posts) divided by the number of words in the submission text. The list of curse words was initially downloaded from bannedwordlist.com prior adding to my own selections.

4.3 Discussion Features

The features in this category focus on measuring both the likelihood of a comment/post generating further discourse and the observed amount of further discussion generated. This is achieved using several feature calculations:

4.3.1 Reply Percentage Score

This score aims to measure the actual amount of discussion generated by a comment or post by essentially calculating the percentage of total discussion activity promoted by a submission. For comments, this is calculated as the count of all children of a comment divided by the total number of comments on the parent thread; For posts, this is calculated as the count of all children comments divided by the total number of comments on the subreddit in the observation period (1 month in this case).

4.3.2 Vote Score

This is simply the score attribute for a given post or comment, retrieved from the Pushshift database. This is calculated as the Lower bound of the Wilson score confidence interval for a Bernoulli parameter.

4.3.3 Link Quantity Score

One way that a comment or post can provide positive discussion value and effectively contribute to discourse is through the inclusion of links. By giving a source for further information, the comment/post simultaneously contributes possibly novel information and provides a focal point for further discussion. This score is calculated for both posts and comments as the count of all links in the submission divided by the number of words in the text.

4.3.4 Question Count

A simple way to increase the likelihood of future discussion generation is by asking questions, as doing so provides a good jumping-off point for future readers to enter the conversation. This score is calculated as the count of questions appearing in the submission text.

Taking a sample of 371 posts and 398 comments, a third-party created annotations by assigning ratings between 1-5 for each with 1 representing a comment with poor direct discussion contribution and poor likelihood of generating further discussion, and 5 representing a comment with high direct discussion contribution quality and a high likelihood of generating further discussion. I then give this rating a binary value of 'Good' for ratings above 2.5 and 'Bad' for those below 2.5, and calculate the feature scores for each comment and post in the data set. Next, I split the sample into a train and test set with a 70/30 split, and train a logistic regression model on the data to predict the binary good/bad value.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
sub_rel	-36.8323	85.0382	-0.4331	0.6649	-203.5041	129.8394
parent_rel	0.5233	0.8354	0.6265	0.5310	-1.1140	2.1607
length	0.1706	0.0616	2.7714	0.0056	0.0500	0.2913
time	-0.3015	0.1062	-2.8381	0.0045	-0.5098	-0.0933
replies	-2.6201	3.3833	-0.7744	0.4387	-9.2512	4.0110
wilson	0.0014	0.0020	0.7150	0.4746	-0.0025	0.0054
links	2.3266	2.5521	0.9116	0.3620	-2.6755	7.3287
qcount	0.4544	0.2482	1.8311	0.0671	-0.0320	0.9409

Figure 1: Comment Logistic Regression Results

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
sub_rel	2.5970	0.7171	3.6216	0.0003	1.1915	4.0025
length	0.1433	0.0625	2.2919	0.0219	0.0208	0.2659
time	-0.0669	0.0273	-2.4534	0.0142	-0.1204	-0.0135
replies	8.1657	21.0848	0.3873	0.6985	-33.1597	49.4911
wilson	-0.0000	0.0000	-0.9819	0.3262	-0.0001	0.0000
links	-0.1896	8.7309	-0.0217	0.9827	-17.3017	16.9226
qcount	-0.1362	0.2503	-0.5442	0.5863	-0.6269	0.3544

Figure 2: Post Logistic Regression Results

5 Evaluation and Results

As the goal of this project was focused on developing a set of features that capture discussion quality on Reddit and other forum sites rather than purely rankings, the main evaluation method for this analysis focuses on using my logistic regression classifiers to measure the predictive power of each of my calculated features. To evaluate my logistic regression classifiers, I use the logistic regression score method and the cross val score from the sklearn library. My comment classifier achieved an accuracy score of 57% and an aggregated 10-fold cross-validation score of 58%, suggesting that the model holds. Likewise, my post classifier model holds as well, achieving an accuracy score of 59% and a 10-fold cross-validation score of 66%. While the results of these regressions could certainly be improved, they are acceptable for use as a baseline for evaluating my scores.

In order to quantify the effectiveness of each of my features in reflecting discussion quality, I compare the beta values of each feature generated by each of my logistic regressions.

Looking at these outputs, we see from the p-values that the majority of my features did not have statistically significant regression coeffi-

cients. For comments, length score and time score were the only significant predictors of discussion quality, although question count is close to significance. For posts, the significant predictor features were subreddit relevance, length score, and time score.

6 Discussion

Given that the goal of this paper was to develop scoring features to quantify discussion quality on Reddit, the results of this analysis lacked the significant implications I had hoped to uncover. The features that I expected to be highly predictive, such as reply percentage and question count, were revealed to be insignificant by my baseline, while the features I sought to make insignificant, mainly timeliness, were shown to be highly predictive. Of the features that I calculated, only the length and time scores were shown to be predictive of whether a submission represented a quality contribution to discussion - a particularly disappointing result considering that a key motivation for this research was the overvaluation of time in Reddit's current scoring algorithm.

The strong predictive power of the time feature was a notably interesting result, as the data provided to third-parties for annotation was not in thread order and did not contain a time feature. It was quite surprising to see that the time feature was significantly predictive for predicting discussion quality, as logic would suggest that without time/order clues to influence annotator rating decisions it would have little correlation with the annotated rating values. However, since the time feature maintained predictive power, this implies that early commenters/posters are more likely to provide quality discussion content - a result that makes intuitive sense, as early submissions have a higher likelihood of accumulating large numbers of replies than later submissions. Beyond the time feature, the length feature was a strong predictor of discussion quality. While these scores are normalized by the submission length, this result still holds, as larger submissions are more likely to generate responses, hold novel information, and contain links and questions.

A strong predictor of post discussion quality - but not comment discussion quality - was relevance to the subreddit topic. Logically, this result follows: the conversational nature of comments in a thread means that a comment can be 'off-topic' in terms of subreddit topic but 'on-topic' for its parent, whereas a posts' role as a high-level thread container provides little leeway for straying from the subreddit topic. In addition to the semantic implications of this result, the significant predictive ability of the post subreddit relevance score indicates strong potential for using LDA topic modeling in a framework such as this. While not ideal for use on small text groupings, like comments, the small p-value and relatively large correlation coefficient illustrate that LDA topic modeling can be an effective way to measure thread and subreddit similarity on Reddit.

Overall, considering the large computation expense required to calculate the features of my algorithm and the lack of significant predictive power of these novel features, this research fails to provide practical ways to improve the current Reddit scoring algorithm - with the possible exception of implementing topic modeling for post scoring. In addition to the poor predictive power of many of my features, this analysis was further hindered by a lack of computational power. I had planned to implement features to measure the originality of a submission's content and the quality of a submission's questions and links, but I was unable to find workarounds with small enough computational expenses. Without these scores, the current regression model fails to properly quantify 'direct discussion quality' as I sought to do.

7 Future Work

An area of weakness in this analysis is the lack of originality measures for posts and comments. I had planned to implement features measuring the novelty of a submissions' content relative to other submissions of the same level (meaning a comment is compared to all previous comments in its parent thread, and a post is compared to all previous posts on the subreddit), but issues with computational expense were too great. If I were able to formulate these scores in the future, I would be able to better explore my analysis' implication that early commenters are more likely to provide quality discussion contributions than later commenters. While this seems like a logical conclusion based on the information at hand, this hypothesis cannot be effectively tested without a way to quantify the actual value of a submission's content. Additionally, I had planned to build features to quantify the quality of links and ques-

tions in posts/comments, but could not implement them due to limited computational power. Without these scores, the ability of my algorithm to capture direct discussion contribution quality is severely hindered. As such, I would certainly focus my energy on these areas in the future.

8 Work Plan

At the beginning of this project, I set out to find a way to effectively quantify discussion quality on Reddit and other, similar forum sites. As my work progressed, the exact format of this analysis shifted from purely focusing on the re-ranking of comments and posts to an approach that focused more on generating human-interpretable insights into the problem of measuring discussion contribution quantity. I originally planned to implement many of these features as the first stage of this research, with the second stage focused more on increasing predictive power by tweaking parameters and exploring additional data mining and machine learning techniques to include in my algorithm. However, I quickly learned that the computations I intended to undertake were much more challenging than I had initially anticipated.

On a positive note, I was happy with the development of my baseline over the course of this project. My original intention was to create a Z-score ranking of comments and posts that would be compared against the actual Reddit score rankings: Yet, I never felt comfortable with this baseline, as it did not really seem to measure the kind of performance I was hoping to produce. As the project progressed, I eventually realized, after talking to Professor Jurgens, that using a classifier to assess the power of my features - a baseline that I felt reflected the goals of my work much more closely.

Knowing what I do now, I would likely have invested more in exploring baseline options from the beginning. A good amount of my project working time was spent without a clear idea of how I would be testing my findings, which negatively impacted the development of the algorithm itself. If I had identified a more relevant baseline at the start of the project, I feel that I would have had a more clearer understanding of the direction I wanted to take.

Acknowledgments

Annotators:
Brandon Punturo
Matthew Jankowski
Josh Hamilton
Jaikishan Prasad
Sean Flanagan

References

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.* 2(CSCW):32:1–32:25. https://doi.org/10.1145/3274301.

Paul Ton. 2017. Topic modeling and user clustering on internet discussion forums - a case study. https://nycdatascience.com/blog/student-works/topic-modeling-user-clustering-internet-discus

Nayer Wanas, Motaz El-Saban, Heba Ashour, and Waleed Ammar. 2008. Automatic scoring of online discussion posts. In *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. ACM, New York, NY, USA, WICOW '08, pages 19–26. https://doi.org/10.1145/1458527.1458534.

Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. 2007. Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 125–128. http://dl.acm.org/citation.cfm?id=1557769.1557806.