Identifying Factors for Asking Effective Questions in Online Forums

Brian Liang
University of Washington
liangb1@uw.edu

Dao Yi
University of Washington
daoyee@uw.edu

Jacob McKenney
University of Washington
jacobmck@uw.edu

Omar Omar
University of Washington
oomar@uw.edu

Vineet Kalki
University of Washington
kalkiv@uw.edu

Abstract

Online forums have provided a platform that has enabled everyone to discuss freely. The limitless possibilities of online forums can make them challenging to navigate, especially in making posts that capture attention and foster productive discussions. Understanding how to use online forums to their fullest extent can be an exceptionally powerful tool, especially in learning and sharing knowledge. However, the unpredictable nature of human interaction makes it challenging to identify the underlying attributes that form a successful post. We conduct a thorough investigation of forum data to identify potential explanatory features behind post interactions. We found that questions that are rated highly, viewed more, and have productive responses have readable titles, sparse tags, and are posted by reputable users. Temporal factors like day of the week and time of day yield varied impact depending on the forum. We come to the conclusion that creating a successful forum post involves making it easy to find, understand, and respond to the post.

1 Introduction

Online forums are a critical tool in our increasingly digitized world as they provide a platform for users to learn from each other, share experiences, and engage with others who have similar interests. However, as these platforms grow, the proportional representation that each user has diminishes, meaning that users may find it increasingly difficult to be heard. This problem can be especially unhealthy in question and answer forums, where a lack of attention will mean that people who need help get no responses. In addition to attracting attention, questionnaires must also get responses that are helpful and productive. This involves the difficult task of motivating a response from someone experienced with the specific problem and framing the question for a good response. With these considerations in mind, posters must meticulously craft the perfect post to get their question answered. Our study aims to address this challenge by answering the question: How can we ask effective questions on online forums?

Answering this question will help us understand forum interactions and harnessing attention. In general, answerers must put in significant effort to answer a question thoroughly but gain no value from this effort other than perhaps minor recognition and a feeling that they helped someone else. Understanding this frequent but puzzling phenomenon will shed insights on what matters to people. Knowing exactly what to do to attract the attention from the right people will simplify the interaction process. Effective interactions will improve discourse, accelerate learning, and enhance knowledge sharing.

However, there are several major challenges to address in this study. Human nature is incredibly unpredictable, which means that the connections between observable features and evaluations in forums can be ambiguous. As a result, naive approaches may not find any meaningful correlations. Additionally, there is a huge amount of variability in forum posts, which can cloud the important differences between successful and unsuccessful posts. This means that even if statistically significant correlations are found, they may not be accurate or impactful.

Previous attempts have been made to address this problem, but they are limited in scope. With a growing forum user base, more data has been produced that allows the problem to be addressed with a more complete data-driven approach. Developments in technology have also improved the ways that this data can be used. Our approach takes advantage of these developments to study this problem completely and thoroughly.

1.1 Sampleset Selection

StackExchange collects all boards on their API under 6 categories: Technology, Culture & Recreation, Life & Arts, Science, Professional, and Business. We aim to represent the largest and most active communities in each of these subsections to form a baseline for all exchange boards. We selected StackOverflow, Mathematics, Quant, Worldbuilding, English, and Workplace as our sample set. The former half focuses heavily on data or technical questions while the latter half better represents experiential and textual questions. Unlike previous studies which focused solely on StackOverflow or trending boards like Bitcoin, this generalized overview extends our studies' external validity and forms a strong basis for StackExchange as a whole (Section 2). For the purposes of narrative simplicity, we will follow the Worldbuilding StackExchange through our criterion and later compare between our selected exchange boards.

The driving blocker for a naive approach is being able to define what comprises a 'good' or 'effective' post. To avoid arbitrarily selecting features to analyze, we surveyed 25 users of StackExchange boards to determine what criteria would best represent their needs. The results dictated that successful questions are answered in a timely manner, are visible in the communities they are posted in, and receive an answer that is accepted. This dynamic of multiple input and output factors led us to devise a strategy to measure and evaluate posts based on patterns in language and trends among exchanges (Section 3). In our analysis, we contribute an evaluation of post composition, post content, and contextual factors (temporal & reputation) measured on average time till acceptance, post visibility, and acceptance rate (Section 4).

1.2 Hypotheses

Furthermore, we hypothesize three possibilities resulting from our findings:
H1. Effective questions rely on esoteric language, detailed content, and many supporting elements.
H2. Effective questions are easy to read, sparse in elements, and concisely written.
NULL. There is no impact from user controlled elements on the effectiveness of a question.

We have found evidence supporting parts of H1 and H2. Questions with 2-3 tags, readable titles, and probing words tend to have more interactions. In general, users should make questions easy for their target audience to view, understand, and answer. Our hypotheses are based on the anonymous nature of online discussion which can follow in extremes to either generate controversy or contain varying levels of detail. Furthermore, the low bar of entry to online discussion and lack of consistent enforcement, other than preventing spam, invites a variety of different perspectives and habits together (Section 5).

2 Dataset

2.1 Dataset Description

The dataset used in this analysis was acquired from the StackExchange public dump [**1**]. It consists of data from six different sites within the StackExchange network: Worldbuilding, English, Mathematics, Quantitative Finance, Workplace, and Stack Overflow. The dataset contains a total of

8,251,569 raw entries, and 3,194,197 entries after filtering. Each entry includes information about one specific question or answer post, and any comments or additional information that was provided by users. There are huge differences in the number of posts across sites. The majority of the posts are from Math and Stack Overflow, which make up 43% and 47% of the dataset, respectively. Our analysis will focus mainly on question posts, which make up 32% of the entries in the dataset. The dataset spans from Jan 2010 to Nov 2022.

2.2 Dataset Preparation

In order to focus our analysis on the most relevant and useful information in the dataset, only question and answer posts are used. We also ignore any posts that do not have an OwnerUserID. Posts without an OwnerUserID are from guest, banned or spam users, and they do not provide valuable information for our analysis.

Missing values are treated carefully in order to ensure that our analysis is accurate and comprehensive. Nullable values that have a meaning (e.g. columns for newer functions that were not available for old data points) are kept. This allows us to accurately represent the data and avoid introducing biases into our analysis. We perform augmentation on the dataset columns in order to add new numeric data from categorical data. Additional insights that might not be immediately apparent from the raw data can be revealed in this way. For example, we convert the tag names into numerical tag counts in order to apply statistical analysis. Instead of normalizing data across sites, we normalize them based on differences in each section.

2.3 Exploratory Data Analysis

We performed univariate and bivariate bivariate analysis on numeric columns.

Univariate analysis will involve studying a single feature or variable in the dataset. This will involve visualizing the data using plots and statistics, such as histograms and density plots, in order to understand its spread and center. Outliers and trends in the data will also be identified.

Bivariate analysis will involve studying the relationship between two variables in the dataset. We will use scatter plots and correlations matrices to identify correlations between different features.

Multivariate analysis studies the relationships between multiple variables simultaneously. However, since there are not many numeric values in the dataset and we get clear trends from univariate and bivariate analysis already, multivariate analysis will not be applied in this section. Instead, it will be performed on the data content analysis in the form of PCA.

2.3.2 Results

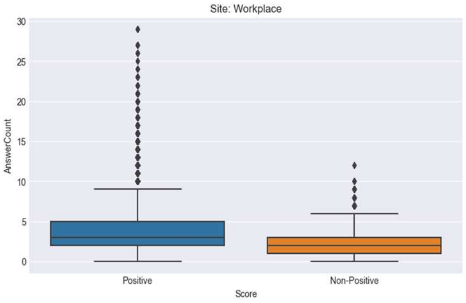


Figure 1: Boxplot of answer counts versus scores on Workplace site. We learn that a post with more than 15 answers on Workplace is likely to have a positive score.

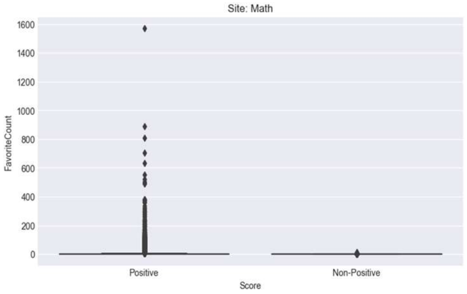


Figure 2: Boxplot of favorite counts versus scores on Math site. It is unlikely for a post with negative score to have any favorites. This implies that favorite counts could be a strong indicator of the quality of a question.

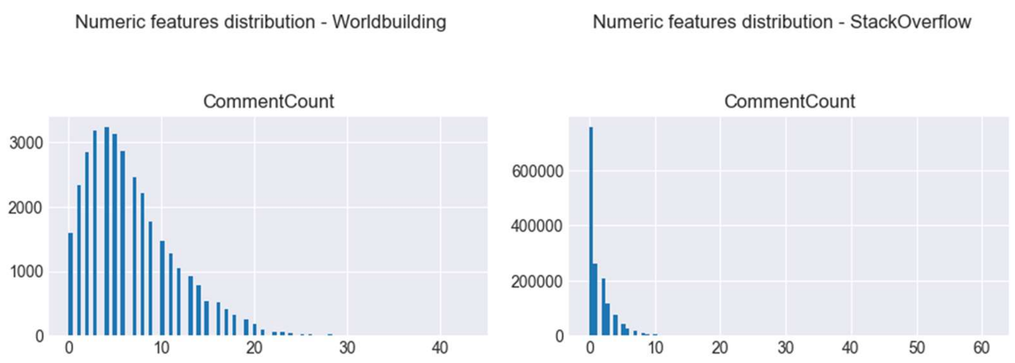


Figure 3: Histograms of comment counts on the Worldbuilding and Stack Overflow sites. Both distributions are bell-shaped, with a clear center and spread.Similar statistical analysis are applying to all variables on all sites. In terms of the mean, 25th, and 75th percentile, we see consistently higher comment and answer counts on Worldbuilding, English, and Workplace than on Math, Quant, and Stack Overflow. The distributions of tag counts are similar across the sites. We suspect this is because sites related to non-STEM disciplines like Worldbuilding and English may require more discussion and commenting by nature. Despite on having least amount of data entries, Quant has largest view counts on all percentiles. We also find that tag counts are consistent on all percentiles across all sites.

By analyzing the trends in values on posts with positive and non-positive scores, we expect to gain insight into what are the distinct characteristics of questions with high scores. Here are a few examples.

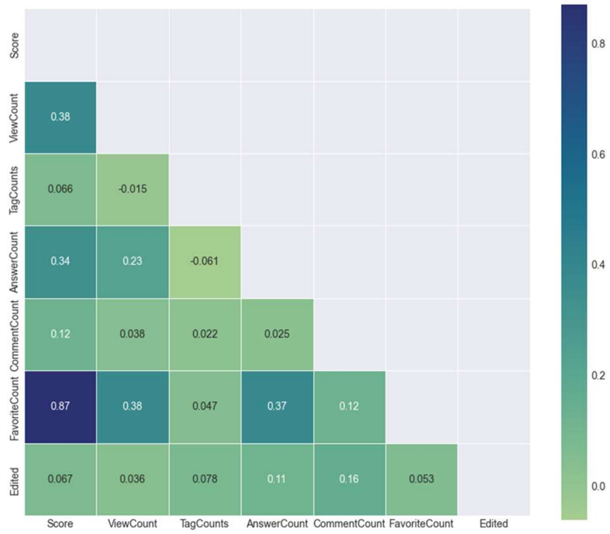


Figure 4: Correlation matrix on variables in Math site. Correlation matrices has similar trends across sites.
High correlation is noticed between Favorite Count and Score which corresponding with previous findings.
Weak correlation is seen between Edited and other variables, and between Answer Counts and Tag Counts

3 Analytical Approach

3.1 Evaluation constructs

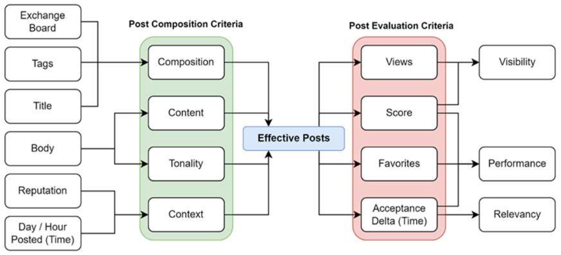
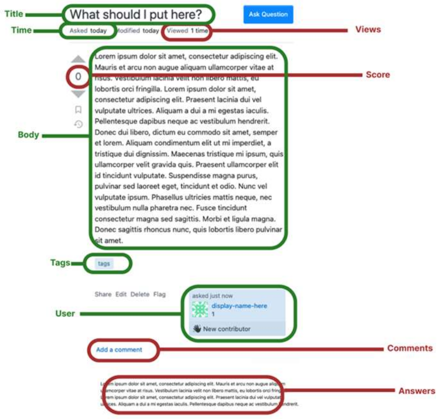


Figure 5: Flow Diagram

A StackExchange post being “effective” can have many different meanings depending on the posting user and their desired outcomes. In this study, we operationalize the construct of post success through four main factors: views, score, acceptance rate, and time elapsed until answer acceptance. For this report we scope these factors down to just acceptance rate and time elapsed until acceptance to allow clearer cross-forum comparison and to align with a survey we conducted that made it apparent getting an answer in a reasonable amount of time is the primary desire for a StackExchange user. A post is composed of many different factors such as title, body text, links, embedded content, time of posting, etc. Drawing on Calefato et. al, we differentiate between four types of factors with the following categories: composition, content, and external factors. Composition is defined as the supporting elements when creating a post; particularly the title and tags attached. Content is concerned with the body text of a post. The former is operationalized through aspects such as complexity of word usage, sentence structure patterns, and uniqueness while the latter is measured via sentiment analysis modeling. Finally, external factors include post attributes such as time of posting and poster accolades.

3.2 Key considerations and limitations

StackExchange forums are rich with information and can give us unaltered access to natural online interactions. However, analyzing these interactions involves processing gigabytes of data, millions of rows of information, and billions of words. This was a huge challenge that we had to mitigate for as we moved into performing data analysis. We addressed this issue through data filtering, pruning down to just the rows we needed, and by improving the compute and memory of the machines we ran our analysis on as well as utilizing distributed analysis frameworks such as Spark for particularly intensive workloads. Additional challenges in analysis included properly parsing body content which we used a popular python library called BeautifulSoup for, ensuring that equations didn't sway our readability scores by editing them out of text portions of our data, and ensuring correctness in our implementations which we alleviated with peer review.



What should I put here?

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris et arcu non augue aliquam ullamcorper vitae at risus. Vestibulum lacinia velit non libero mattis, eu lobortis orci fringilla. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent lacinia dui vel vulputate ultrices. Aliquam a dui a mi egestas iaculis. Pellentesque dapibus neque ac vestibulum hendrerit. Donec dui libero, dictum eu commodo sit amet, semper et lorem. Aliquam condimentum elit ut mi imperdiet, a tristique dui dignissim. Maecenas tristique mi ipsum, quis ullamcorper velit gravida quis. Praesent ullamcorper elit id tincidunt vulputate. Suspendisse magna purus, pulvinar sed laoreet eget, tincidunt et odio. Nunc vel vulputate ipsum. Phasellus ultricies mattis neque, nec vestibulum nulla pharetra nec. Fusce tincidunt consectetur magna sed sagittis. Morbi et ligula magna. Donec sagittis rhoncus nunc, quis lobortis libero pulvinar sit amet.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris et arcu non augue aliquam ullamcorper vitae at risus. Vestibulum lacinia velit non libero mattis, eu lobortis orci fringilla. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent lacinia dui vel vulputate ultrices. Aliquam a dui a mi egestas iaculis. Pellentesque dapibus neque ac vestibulum hendrerit.

Figure 6: Element by element breakdown of post composition criteria (green) and post evaluation criteria (red).

3.3 Validity

We were careful not to introduce bias into our study which was bolstered by the fact that the data used from the StackExchange archive was complete, containing all data from the creation of a forum until September 2022. This meant no bias was introduced in data collection as we have a dataset created quite randomly by natural user interaction. Additionally, our data filtering to remove posts that had been edited helped us stay internally consistent as it helped us remove bias in the relationship between success metrics to the actual content that prompted that success. Our study is externally valid because we took data straight from the source we were studying and looked through the entire history of posting, this allows us to make the claim that our findings can be applied actionably to the same forum we find them in for future success within that context. Moreover, our findings can be generalized easily to other stack exchange forums that weren't analyzed in our study as the StackExchange archive exposes the exact same set of tables with the same set of features for every single forum. This makes it not only feasible, but quite easy to perform our same analyses on more forums with only minor changes to account for forum-specific content differences like code blocks in StackOverflow.

4 Results and Findings

4.1 Composition

Composition encompasses the supporting elements of a post that users have control over from its inception. Outside of the body and selecting the question's domain, users are able to set a 150 character title and select between one to five tags. Titles and tags are typically the first phase of exposure to questions on StackExchange boards. Indexed results, search queries, and categorical sorts rely on word matches or filtering using the data from these two elements.

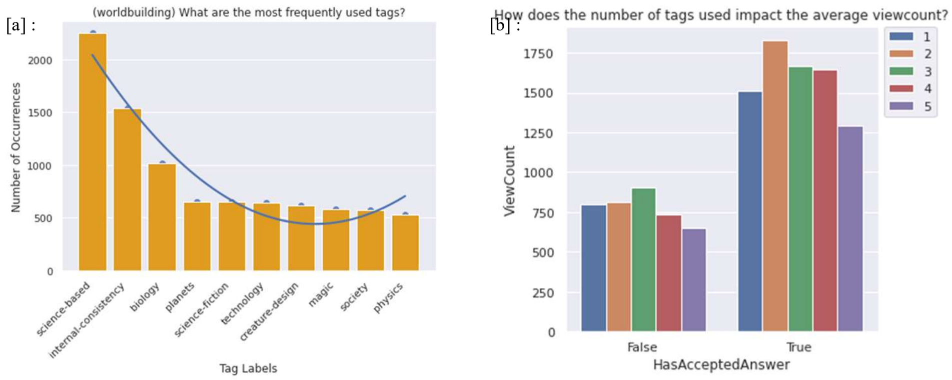


Figure 7: Breakdown of tag usage in the Worldbuilding Exchange. [a] depicts heavy favorability to a limited number of tags while [b] illustrates how 2-3 tags garner faster responses than others.

When comparing the quantity of tags used to the average views a post gets, we can see a trend favoring frugality. Posts that use either two or three tags saw a 21.27% increase in mean view count over alternate setups. This indicates that there is stronger organizational value in selectively applying tags in the Worldbuilding exchange. Furthermore, tag usage is highly concentrated around three tags: science-based, internal-consistency, and biology. This subset accounts for 16.36% of 29,381 tagged instances. Users are more likely to encounter posts when filtering by the most commonly applied tags. The share held by the top three tags parallels the conclusion that using 2-3 tags will maximize visibility since using more is unlikely to have as substantial of an impact.

To measure title readability, we applied the Flesch-Kincaid reading ease test which measures the perceived difficulty of an English passage. It uses a combination of lexicon count, average syllables per word, and sentence length. To see how title readability correlates to the average time to acceptance, we compared the reading score in increments of twenty-five to the average hourly delta between when a question was asked and when an answer was selected as 'accepted'. When looking at the Worldbuilding exchange, we see a clear inverse relationship (r-squared of -0.93) between these two factors. Titles at the lowest band of [-25, 0) averaged an acceptance delta of 4.497 hours whereas those on the higher extreme averaged 3.656 hours. This indicates that the worldbuilding community is more likely to answer questions quickly when titles are easier to read which falls in line with our initial expectations.

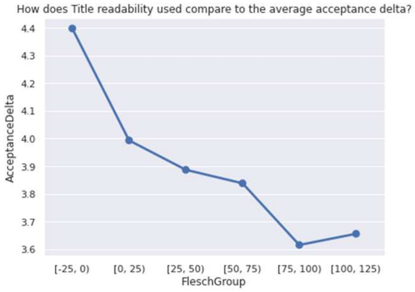
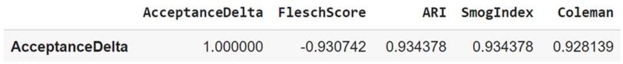


Figure 8: As Flesch scores increase, Worldbuilding sees a decrease in the average hours required to find an accepted answer.



	AcceptanceDelta	FleschScore	ARI	SmogIndex	Coleman
AcceptanceDelta	1.000000	-0.930742	0.934378	0.934378	0.928139

Figure 9: R_Squared breakdown of various readability tests versus time till acceptance delta in hours.

This result is further supported when we look at similar readability tests such as the Smog Index, Automated Readability Test, and the Coleman Liau Index. Similarly to the Flesch-Kincaid test, these approaches use textual data to measure and grade the difficulty of a passage. However, these indexes rely on the traditional American grade level scale system where a score of 9 would be equivalent to a 9th grade reading level. Thus the strong positive correlation observed also indicates that the difficulty of a title is proportional to increased wait times.

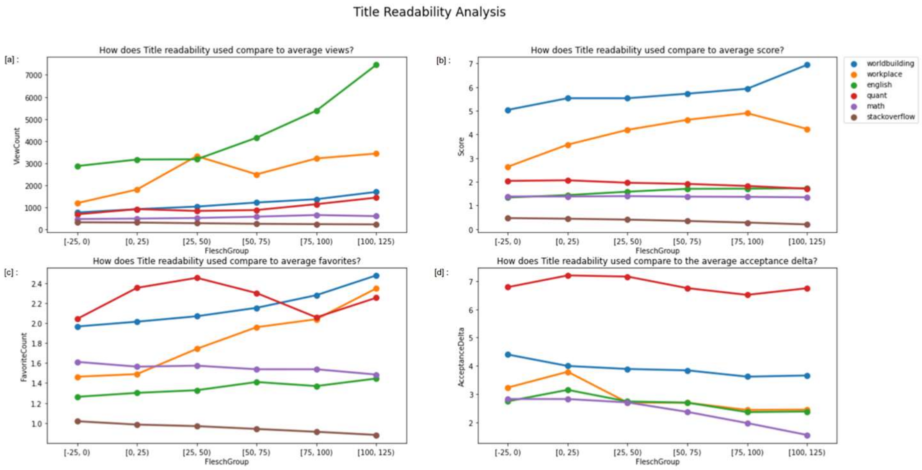


Figure 10: Multi-faceted breakdown of the impact of title readability across discussion boards. [a, b, c] illustrate the direct relationship between visibility factors and text based exchanges as well as the inverse relationship to data based exchanges. [d] illustrates how regardless of forum, the acceptance delta decreases with better title readability.

When taking a look at how title readability reflects across other forums, we generally see an agreement that improved readability tends to positively impact factors like average view count, score, favorite count, and time till acceptance. However, looking closely we notice that data-driven forums such as StackOverflow and Math trend opposite for visibility factors (views, favorites, score). A potential reason for this is a heavier reliance on LaTeK equations in titles as supporting elements which is scored negatively by readability algorithms. These exchanges tend to rely on examples and information which supports the argument that textual based exchanges follow directly to readability while data heavy exchanges are more likely to be inversely related.

4.2 Content

To analyze the content of the post, we investigate wording and sentence structures of the post title and body. As a general post set, we use a set of 1,000 posts sampled randomly from the 148 forums on StackExchange with over 1,000 valid posts to have a balanced post set encompassing all reasonably sized StackExchange forums.

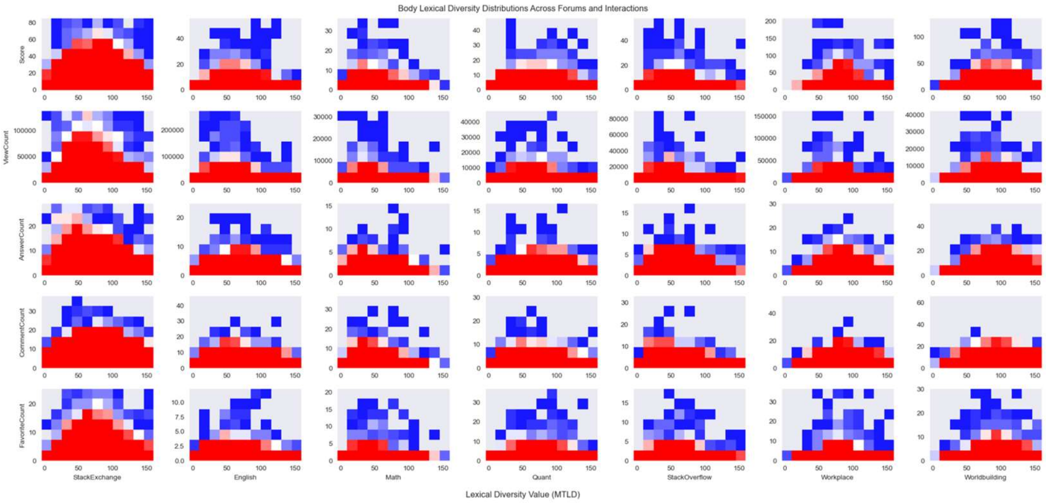


Figure 11: Density scatter plot of post body lexical diversity values and interaction quantities show interaction quantities peak with lexical diversity values of 50-100 across all forums.

We analyzed the impact of specific words on interaction quantity by calculating the differences in interaction quantities between posts titles that contained and did not contain the question tokens: {who, what, where, when, why, how, '?'}. We found that on average, the inclusion of question tokens increased interaction quantities in almost every category by fairly significant margins. Question marks, which appeared in half of post titles, were correlated with an acceptance rate increase by 6.8% and view count increase by 932 on average (Figure). Across forums, results varied, but a general increase in interaction quantities was still observed. We analyzed the impact of word choice by observing the patterns between interaction quantities and Measure of Textual Lexical Diversity (MTLD) values, a formula that tracks token uniqueness through token frequencies and number of tokens [4]. We found that across all forums, interaction quantities peaked between MTLD values of 50-100 (Figure). This observation shows that post bodies should have not too many repetitions, but not too few repetitions either, and that using a balanced vocabulary should yield higher interaction quantities.

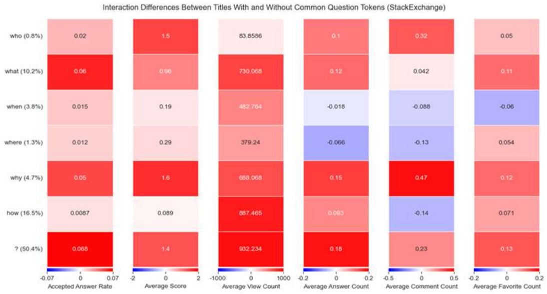


Figure 12: The average interaction quantities of posts with titles that include question tokens are higher than those of posts with titles without question tokens.

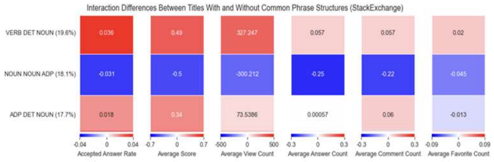


Figure 13: The average interaction quantities of posts with titles that include the structures {verb, determiner, noun} and {adposition, determiner, noun} are higher than those of posts with titles that do not include the structures.

We analyzed the impact of sentence structures on interaction quantity by identifying the common parts of speech structures across posts, and measuring the differences in average interaction quantities between post titles that contained and did not contain these structures. The two recurring structures that appeared frequently in the general set and in the individual forums were the structures {verb determiner noun} and {adposition determiner noun}. Notably, in the general set, post titles with these structures saw a 3.6% and 1.8% acceptance rate increase and 327 and 73 more views on average, respectively (Figure). Quantities varied, but a general increase was observed across boards as well.

4.3 Contextual factors

In addition to the composition and content of a post there are other contextual factors that are important to consider when evaluating post answerability. Three of these factors that we analyzed during our research are time of posting, reputation of poster, and badges associated with poster.

By looking at how our post evaluation criteria metrics vary over the week throughout the course of the entire history of the Worldbuilding forum we can begin to see some themes and thus provide users with some Worldbuilding-specific information to help them choose when it is that they should prioritize posting. The graph below shows that acceptance rate and mean hours until acceptance are optimized (high and low respectively) in the afternoon and especially on weekdays. We see a drop in acceptance rate and increase in hours until acceptance on Friday's and then a continual return to weekday standards through the weekend. This shows a clear preference regarding acceptance rate and timeliness weekday afternoons for the Worldbuilding forum.

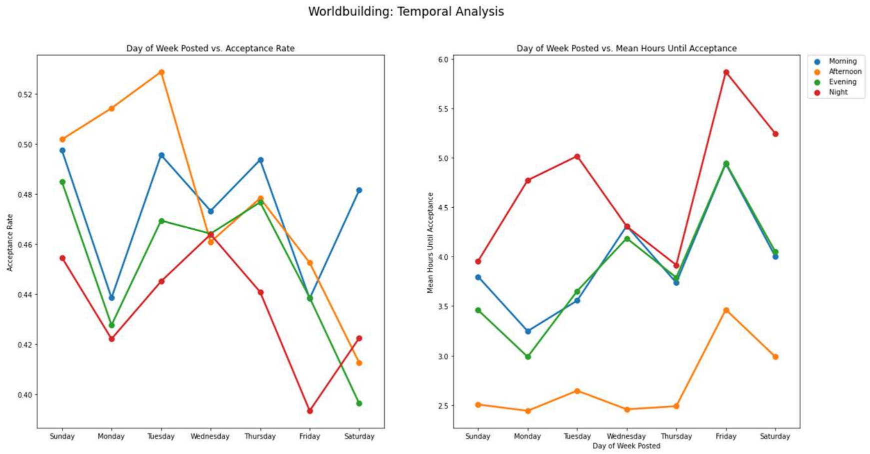


Figure 14: Worldbuilding maximizes acceptance rate and minimizes acceptance punctuality during afternoons on weekdays

On StackExchange sites users can build reputation through various means but primarily through posts being upvoted and accepted. This means that user reputation is a fairly good proxy for post quality as it is a crowd sourced estimate of a user's domain knowledge for a given forum. To see how reputation correlates with acceptance rate we graphed the top x percent of equations by poster reputability to the corresponding acceptance rate for those questions and found a significant negative correlation (r-squared of 0.46 and p-value < 0.05). This gives us insight into the importance of establishing credibility within a community such that you are trusted. Prominent figures in communities get more interaction and have more people following them closely. Unsurprisingly, there is a positive correlation, albeit weaker (r-squared of 0.22, and p-value of 0.19), when comparing the same question reputation groups (filtering out non-accepted answers) to their mean time until acceptance. This shows a similar trend to the temporal analysis where acceptance rate and time to acceptance tend to be inversely related to each other in some measure.

Another StackExchange construct in the same vein as reputation is badges. Badges are awarded to users for a wide variety of reasons, all of which involve time and commitment invested into the forum. Some are general such as "Socratic" which is awarded upon asking 100 well-received questions on distinct days and others are tag-specific such as "Science-based" which reflects a continued investment and expertise in that category. For Worldbuilding we found the top-5 badges by acceptance rate of users who hold those badges are Universe, Crime, Mathematics, Time, and Climate-Change. Additionally, using the same analytical approach as user reputation we compared top x percent of users by number of badges against both acceptance rate and mean hours until acceptance and found a very similar relationship as user reputation. This strengthens our findings of inverse relation between acceptance rate and mean hours until acceptance as well as implies user reputation and number of badges are positively correlated.

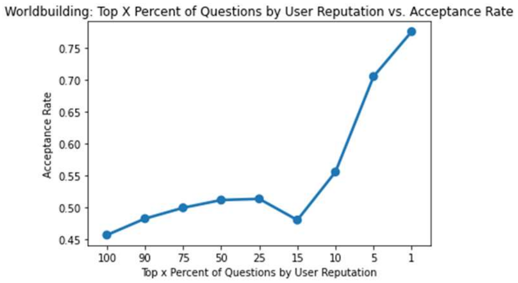


Figure 15: Establishing reputability in a community is key to post engagement

When expanding our analysis to multiple forums in regards to time of posting we find that worldbuilding, workplace, and english which are the more text-oriented forums, have noticeably higher acceptance rates during the weekdays (an average 5% increase over weekends) during afternoon whereas the other three equation-preferential forums tend to nominally higher acceptance rates during evening and night but still higher response times during those hours like the text-oriented forums. Additionally, these equation-based forums have substantially higher acceptance rates (up to 3% higher) on weekends.

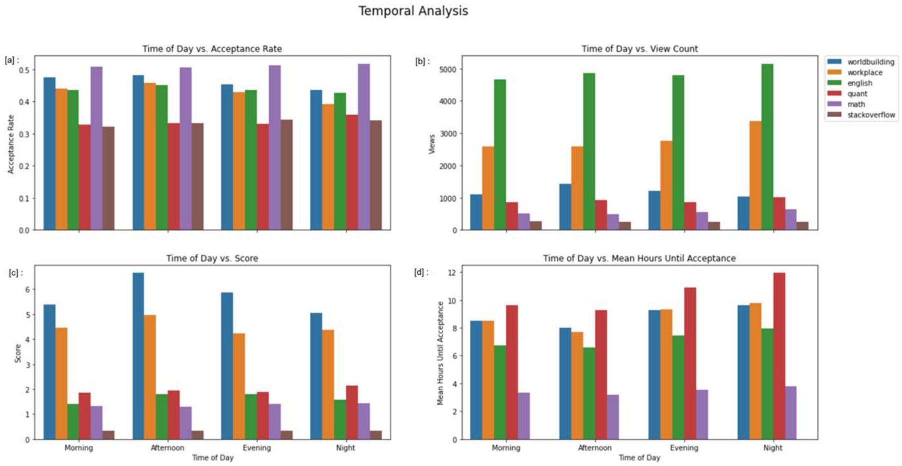


Figure 16: By comparing themes throughout our six forums we see trends in text-based and equation-preferential forums in acceptance rate [a], view count [b], score [c], and mean hours until acceptance [d]

5 Related Work

Calefato et al. have conducted similar studies on how characteristics of questions on StackExchange play into the success of the question. One study analyzed the impact of affect, presentation quality, time, and reputation on whether or not a question is answered on StackOverflow. Their conclusions in this study found that reputation has the greatest impact on question success in StackOverflow, but actionable factors include making questions short, including code snippets, using standard casing, and adopting a neutral emotional style [3]. Our study differs in that we study a broader range of question forums and features, allowing for greater generalizability. Another study analyzed the impact of affect, presentation quality, and time on the success of a question in the arcade, bitcoin, math, and science fiction StackExchange forums. The conclusion was that presentation quality has the most significant impact, with day and time also having a measurable impact [2]. Our study is more thorough as we consider more forums and features.

6 Conclusion

We have presented a system for evaluating the effectiveness of questions online by judging the composition, content, and context of a post against the time delta to the acceptance of a post and the acceptance rate of posts across StackExchange forums. Our findings supported a combination of our hypotheses as follows. We discovered that readable titles and the sparse use of tags yield quicker and higher concentrations of results from the composition standpoint. The content of the body consistently has shown little to no influence over acceptance rate or time deltas which contrasts the findings of other sections. Finally, different forums have different optimal times of day to garner responses but engaging in those communities to build reputation consistently improves acceptance rates. For future studies, our criterion can serve as a template to build on when conducting in-the-field research instead of passive observation. The selective modification of the traits we outlined we further prove or disprove our findings and ultimately our hypotheses.

7 Referenced Works / Citations

[1] Stack Exchange Data Dump : Stack Exchange, inc. : Free Download, borrow, and streaming. Internet Archive. (n.d.). Retrieved December 11, 2022, from https://archive.org/details/stackexchange

[2] Calefato, Fabio, et al. "How to Ask for Technical Help? Evidence-Based Guidelines for Writing Questions on Stack Overflow." 2018.

[3] Calefato, Fabio, et al. "How to Ask for Technical Help? Evidence-Based Guidelines for Writing Questions on Stack Overflow." *Information and Software Technology*, vol. 94, 2018, pp. 186–207., https://doi.org/10.1016/j.infsof.2017.10.009.

[4] McCarthy, P.M., Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42, 381–392 (2010). https://doi.org/10.3758/BRM.42.2.381