Min(e)d Your Tags: Analysis of Question Response Time in StackOverflow

Vasudev Bhat Adheesh Gokhale Ravi Jadhav Jagat Pudipeddi Leman Akoglu
Department of Computer Science
Stony Brook University
Email: {vabhat, agokhale, rjadhav, jpudipeddi, leman}@cs.stonybrook.edu

***Abstract*—Given a newly posted question on a Question and Answer (Q&A) site, how long will it take until an answer is received? Does response time relate to factors about how the question asker composes their question? If so, what are those factors? With advances in social media and the Web, Q&A sites have become a major source of information for Internet users. Response time of a question is an important aspect in these sites as it is associated with the users' satisfaction and engagement, and thus the lifespan of these online communities. In this paper we study and estimate response time for questions in StackOverflow, a popular online Q&A forum where software developers post and answer questions related to programming. We analyze a long list of factors in the data and identify those that have clear relation with response time. Our key finding is that *tag-related* factors, such as their "popularity" (how often the tag is used) and the number of their "subscribers" (how many users can answer questions containing the tag), provide much stronger evidence than factors not related to tags. Finally, we learn models using the identified evidential features for predicting the response time of questions, which also demonstrate the significance of tags chosen by the question asker.**

***Keywords*-online communities, question answering sites, collective intelligence, question response time, user engagement, human behavior, evidential feature analysis**

I. Introduction

Q&A sites like StackOverflow, Yahoo! Answers, Naver, Quora, LiveQnA, WikiAnswers etc. are becoming increasingly popular with the growth of the Web. These are large collaborative production and social computing platforms of the Web, aimed at crowd-sourcing knowledge by allowing users to post and answer questions. They not only provide a platform for experts to share their knowledge and get identified but also help novice users solve their problems effectively. StackOverflow[1] is one such community-driven Q&A website used by more than a million software developers who post and answer questions related to computer programming. It is governed by a reputation system[2] which rewards the users by giving reputation points, badges, extra privileges on the website, etc. by the usefulness of their posts. The usefulness of a question or an answer is largely determined by the number of votes it receives.

In such a crowd-sourced system driven by a reputation mechanism, response time of questions to receive the first answer plays an important role and would largely determine

[1]http://www.stackoverflow.com
[2]http://meta.stackoverflow.com/help/whats-reputation

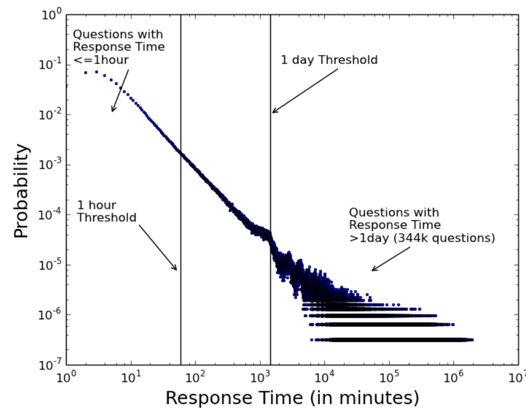


Fig. 1. Probability distribution of the response time of questions on StackOverflow.

the popularity of the website. People who post questions would want to know the time by which they can expect a response to their question. In fact, a study by Rechavi *et al.* [1] on the response time and satisfaction rates in Yahoo! Answers showed that askers use the response time as a measure for marking the best answer, which indicates that the askers are looking for quick responses. Response time analysis would also help improve the site where administrators can reach out to the developers who possess necessarily skills so that the questions get answered quickly. Although most questions on StackOverflow are answered in less than an hour, we observe that about 30% of the questions which are not answered within an hour have a response time of more than a day (see Figure 1). Therefore, it remains a crucial task to infer the response time of questions in Q&A sites in order to help improve user satisfaction and lifespan.

Related to response time analysis, Mahmud *et al.* [2] proposed models based on exponentially-distributed wait times to predict response time of questions on Twitter. Avrahami *et al.* [3] analyzed response times in instant messaging based on the IM app and desktop settings. Our response time analysis on StackOverflow is similar, however, a number of substantially different factors such as the reputation mechanism, tag information, availability of experts etc. are involved in StackOverflow which influence response time. Prediction of response time in a Q&A site is thus a challenging problem due to the number of factors involved.

In our work, we identify evidential features related with question response time and use these factors to build prediction models. As our key contribution, we show that besides several other factors, tags of a question have strong correlation with and are extremely indicative of response time. Tagging questions involves askers selecting appropriate keywords (e.g., `android, jquery, c#`) to broadly identify the domains to which their questions are related. There also exist mechanisms by which other users can subscribe to tags, search via tags, mark tags as favorites, etc. As a result, tags play a crucial role in how the questions are answered and hence determining their response time. We focus on various tag-related features, such as the frequency of tags and number of subscribers of tags, which aid us in building our prediction models. To the best of our knowledge, none of earlier works in the literature have studied the tag-related features and their influence on response time in Q&A sites.

Following are the main contributions of our work:

- We study a large set of factors likely to be associated with question response time in question answering sites. For the first time in the literature, we analyze *tag-based* features and illustrate the strong correlation between question tags and response time.
- We propose to exploit tag-based features in estimating response time on a large collection from StackOverflow. Experiments demonstrate the utility of tags in the prediction tasks.

II. Q&A Sites and Data Details

Formal articles and books are often inadequate in providing answers to questions that people have on a daily basis. Several Q&A sites have become popular that meet the needs of Internet users in seeking answers to their questions. Morris *et al.* [4] studied the kinds of topics for which people turn to their social network to seek answers about and found *technology* to be the top contender.

StackOverflow is a popular community-driven *technology-focused* Q&A site, which is used by more than a million developers across the globe who post and answer questions related to computer programming. Example questions include *"How to send 16bit data to FPGA??"*, and *"How to specify file order with Grunt?"*. Askers can also specify a maximum of 5 keywords, i.e. tags, that broadly describe the domain which their questions belong to. For example, the latter question above on Grunt contains three tags, `node.js, gruntjs, minimatch`.

Users of StackOverflow can either post answers to questions or comment on them, asking for more details. Questions and answers can be up-voted or down-voted, deleted, or edited by the user or the site moderators. People who answer and post questions are rewarded via the reputation system[3] which rewards the users by giving reputation points and these points depend on the number of votes the question or the answer receives. Other rewards include badges for users, bounty (i.e., a lump sum reputation transfer from one user to the other) for answering specific questions, extra privileges on the website, and so on.

[3]http://meta.stackoverflow.com/help/whats-reputation

StackOverflow provides a public dump of its database in every quarter. For our analysis we consider data of four years which spans from July 31, 2008 to July 31, 2012. This data has information about all the posts, users, and votes. The size of the data in total is approximately 17GB. Several statistics of the data are provided in Table I.

TABLE I
STACKOVERFLOW DATA STATISTICS.

Users: 1.3 million
Questions: 3.4 million, Answers: 6.8 million
Questions answered: 91.3%
Median time to receive an answer: 16 minutes
Questions answered in $\leq$ 1hr: 63.5%
Questions answered in > 1 day: 9.98%
Expected number of tags a question has: 2.935

III. Research Questions

Our goal is to investigate which features of StackOverflow are highly correlated with response time and to build a model that can effectively estimate the response time of questions. More formally, we try to address the following problem: Given a set of questions $Q_1, Q_2, Q_3, ...Q_n$ and the response time of their first answer $R_1, R_2, R_3, R_4, ...R_n$, where R_1 is the response time of the first answer received for question Q_1, R_2 is the response time of the first answer received for question Q_2 and so on, predict the response time R of a new question Q that has been asked by a user, which is yet to receive an answer.

Specifically, we formulate the response time prediction problem as two separate classification tasks:

Task 1.
Given a question (its tags, body, title, post date),
Predict if it will be answered in ≤ 16 minutes (median response time) or not.

Task 2.
Given a question (its tags, body, title, post date),
Predict if it will be answered in ≤ 1 hr or > 1 day

We start with a conjecture: The asker-specified tags of a question have significant influence on its response time, since users often answer questions based on their broad domains as specified by the tags. For example a question with tag `android` is likely to attract a specific group of answerers, which may be different from one with tag `genetic-programming`. Tags are important, because users can subscribe to and follow certain tags, search the site based on tags, and designate certain tags as their favorites, which help them quickly identify the questions of relevance to their expertise and interest. The list of research questions we are interested in answering through this study are listed as follows:

1) What are the intrinsic factors and signals that are likely to influence a question's response time?
2) What site-level information is available that shows significant correlation with response time? How do tag-related factors relate to response time?
3) Can we predict question response times using the evidential features available on the site? How effective are the tag-based features?

IV. Data Analysis

In this section we first describe the evidential features we considered and found to correlate with response time. Later, we show several results of our data analysis demonstrating these correlations.

A. Evidential Features

We construct and study a long list of potentially indicative features in estimating the response time of questions. We group our features into two: those that are tag-related and those that are not. We give a list of all the features and their short descriptions in Table II.

TABLE II
FEATURES CORRELATED WITH RESPONSE TIME. WE GROUP FEATURES INTO TWO: NON-TAG BASED FEATURES AND TAG BASED FEATURES. NON-TAG BASED FEATURES ARE STUDIED EARLIER WHILE WE ARE THE FIRST TO PROPOSE AND STUDY TAG BASED FEATURES FOR RESPONSE TIME PREDICTION.

Tag based Question Features
tag_popularity: Average frequency of tags
num_pop_tags: Number of popular tags
tag_specificity: Average co-occurrence rate of tags
num_subs_ans: Number of active subscribers
percent_subs_ans: % of active subscribers
num_subs_t: Number of responsive subscribers
percent_subs_t: % of responsive subscribers
Non-tag based Question Features
num_code_snippet: Number of code segments
code_len: Total code length (in chars)
num_image: Number of images
body_len: Total body length (in chars)
title_len: Title length (in chars)
end_que_mark: Whether title ends with question mark
begin_que_word: Whether title starts with 'wh' word
is_weekend: Whether question posted on weekend
num_active_verb: Number of verbs that indicate action
num_selfref: Number of self references of the asker

1) Tag based Question Features: The tag based features are the main contributions of this work, since these have not been studied in any of the earlier works on Q&A sites, let alone for predicting response time.

tag_popularity: We define popularity of a tag t as its frequency, i.e., the number of questions that contains t as one of its tags. For each question, we then compute the average popularity of all its tags.

num_pop_tags: Each question can contain a maximum of 5 tags. Here, we set a threshold on the frequency of tags to group them into popular and non-popular ones, and count the number of popular tags each question contains. We derived three such features based on frequency thresholds $25, 50$, and 100.

tag_specificity: We define the "togetherness" of two tags as the extent to which the two co-occur in a question and we measure it using the Point-wise Mutual Information: $PMI(x,y) = \frac{p(x,y)}{p(x)p(y)}$, where $p(x,y)$ is the probability of tag x and tag y occurring together in a question and $p(x)$ is the probability of tag x occurring in a question. The specificity of a question is the average "togetherness" of all pairs of tags that it contains.

In the following, we describe the features related to the "subscribers" of tags. Subscribers of a tag are defined as those users who usually answer questions containing that particular tag. Our goal is to quantify the number of "active" and "responsive" subscribers for each tag. The activeness is associated with the amount of questions with a certain tag that a user is capable of answering. The responsiveness is associated with the speed with which the user answers questions containing a certain tag. As such, we calculate the number of subscribers of a tag t based on (1) the number of answers posted by a user to questions containing t and (2) the user's average response time to questions containing t.

num_subs_ans: We define an "active subscriber" of a tag t to be a user who has posted "sufficient" answers in the "recent past" to questions containing t. We say that a user has posted "sufficient" answers when the number of their answers is greater than a particular threshold $\delta(ans)$ and by "recent past" we mean a predefined number of months $\delta(mo)$ before the last posted answer in the dataset. We conducted experiments with $\delta(mo)$=3 and $\delta(ans)$ = $10, 20, 30$. After computing the number of active subscribers for every tag, we compute the average number of active subscribers for individual tags in each question.

percent_subs_ans: We also compute the ratio of the number of "active subscribers" to the total number of subscribers, where the total number of subscribers indicates the number of users who have posted at least one answer ("in the recent past") to a question containing a particular tag.

num_subs_t: We say that a user is a "responsive subscriber" of a tag t if their average response time for questions containing t and posted in "recent past" is less than a threshold $\delta(t)$. We set $\delta(mo)$=3 for defining recent past as before, and $\delta(t)$=1 hour. We then average the number of responsive subscribers of the individual tags each question contains.

percent_subs_t: Similarly, we also compute the ratio of the number of "responsive subscribers" to the total number of subscribers, where the total is defined as before.

2) Non-tag based Question Features: The non-tag based features are quite straightforward and their short descriptions in Table II are explanatory enough so we do not discuss them in detail here. The two that deserve a longer description are the following.

num_active_verb: Active verbs are those that indicate certain action taken by the asker before posting the question which (s)he mentions in the description of their question. Examples include “tried”, “did”, “made”, “used”, “run”, etc.

num_selfref: Self references are words such as “I”, “we”, “me”, “my”, “myself”, etc. which the asker uses to refer to himself/herself or his/her work.

The above two features indicate that the user has done certain amount of ground work before asking the question. It is likely that such words implying prior effort increase the strength of the question, which thereby receive an early response. Both features are normalized by the question length.

B. Feature Analysis

To analyze the question features and their correlation with response time, we constructed two types of plots: (i) box plots (feature value against the response time) that capture the median, 25% and 75% of the distributions, as well as the minimum and maximum values, and (ii) cumulative distribution function (CDF) plots of the response time. We bin the values of most of the features using (base 2) logarithmic binning [5] except for the features considering ratios since the range of percentage is between 1-100.

1) Tag based Question Features: In Fig. 2 we observe that as the popularity of tags increases the response time decreases, which implies that using more popular tags in a question is likely to result in an earlier answer on average. Similarly, the response time drops with the count of most popular tags that the question contains. On the right figure, we note the significant difference in response time among questions with at least one popular tag versus those that contain *none*.

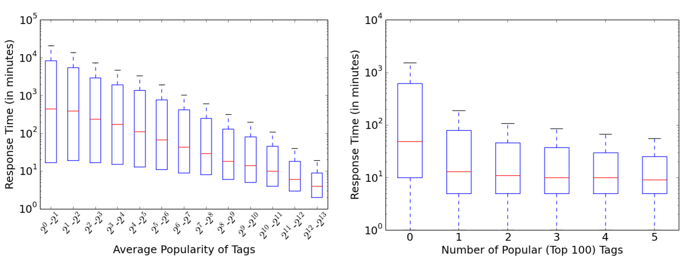


Fig. 2. Response time drops by increasing popularity of tags and number of popular tags.

Fig. 3 show that the specificity of the questions is positively correlated with response time. This can be attributed to the fact that if a question has tags that are too specific, then it is less likely that an answerer would find this question easily and hence the response time of the question increases.

In both types of subscribers (“active” and “responsive”) we observe that the response time decreases as the number of subscribers increases. In Fig. 4 we show the cumulative distribution of response time for various ranges of number of such subscribers. Specifically, we see that the probability of a question having a response time, e.g., less than 1 hr, is greater in the case of questions with higher number of subscribers when compared to questions with fewer subscribers.

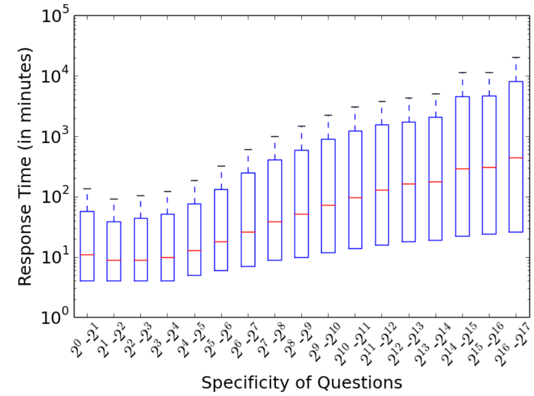
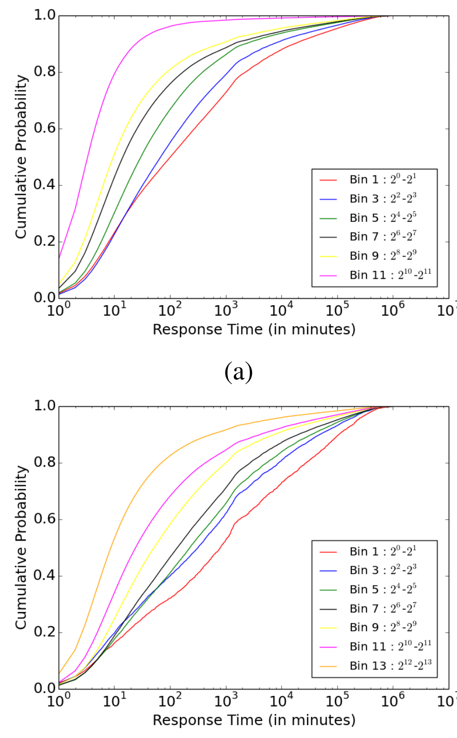


Fig. 3. Response time increases by the specificity of tags a question contains.



(a)

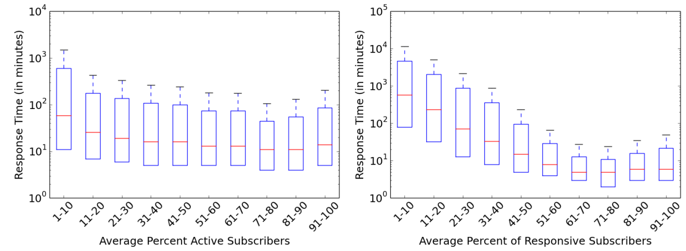
(b)

Fig. 4. Cumulative probability distribution of the response time for *number* of (a) active and (b) responsive subscribers. Higher bin numbers correspond to larger counts and lower response time.

Similarly, the larger percentage of active and responsive subscribers also correspond to lower response times as shown in Fig. 5. As expected, response time drops more with more responsive subscribers than with more active ones, as the former is directly associated with their answering speed.

In summary, all of our tag based features prove to be strongly correlated with question response time.

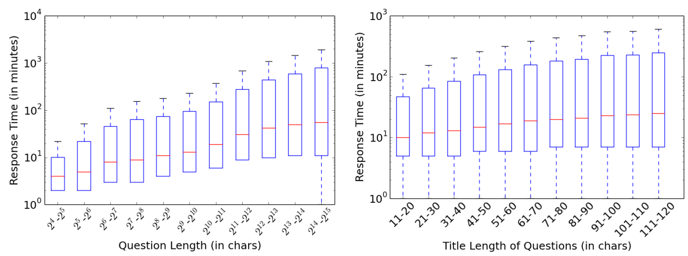
2) Non-tag based Question Features: Next we analyze the correlation of non-tag features. Fig. 6 shows that as both the post length as well as questions’ title lengths increase, re-



(a) (b)

Fig. 5. Response time decreases for larger *percent* of (a) active and (b) responsive subscribers.

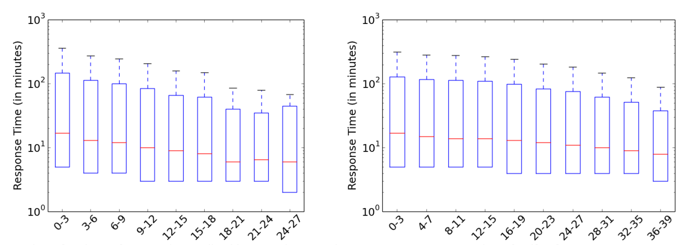
sponse time also increases. In other words, succinct questions seem to receive faster responses.



(a) (b)

Fig. 6. Response time increases by questions' (a) body length and (b) title length.

With action verbs and self references, we see in Fig. 7 that response time slightly decreases as the number of these words increases although the decrease is not as significant as we observed for tag-related features such as those based on tag popularity and subscriber counts. As such, these are likely to provide weak evidence for estimating response time.



(a) (b)

Fig. 7. Response time drops only slightly with the number of (a) action verbs and (b) self-references.

Fig. 8 (a) shows that including code snippets in questions is likely to help with faster response however only if the codes are of certain length (∼50-150 characters). Too short or increasingly longer codes tend to increase response time. Finally, Fig. 8 (b) shows that questions that contain "?" or/and start with "wh" words in their title are more likely to receive faster responses, although the drop in response time is quite small and thus they also seem to be weak indicators.

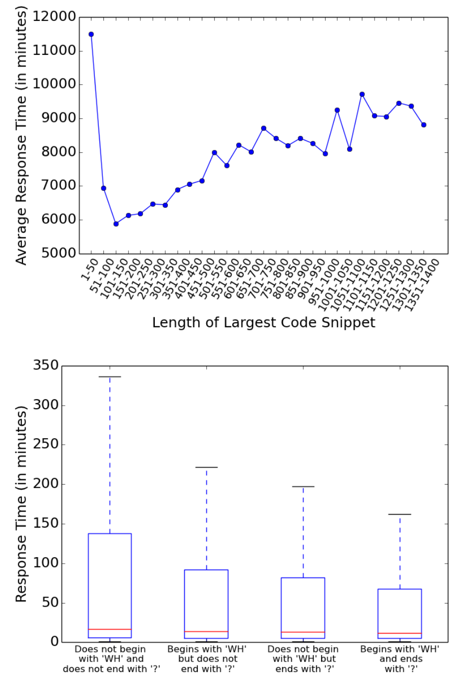


Fig. 8. Response time (top) by largest code length embedded in the questions, and (bottom) based on whether the question title is framed as a question or not.

3) Feature Analysis Summary: For brevity, we present and discuss only a subset of the feature analysis figures. In the following, we summarize our observations of all features that positively and negatively correlate with response time.

Observation 1 (Positive Correlations): Given a question, the higher its body length, title length, code length, and tag specificity; the higher its response time tends to be.

Observation 2 (Negative Correlations): Given a question, the larger its tag popularity, number of popular tags, number and percentage of active and responsive subscribers associated with its tags, its number of active verbs and self-referencing words, number of code snippets (up to 5), number of images (up to 5); the lower its response time tends to be. Starting the title with a "wh" word and ending it with a question mark as well as posting the question on weekend days also correlate with lower response time.

V. Response Time Prediction

Having identified a long list of evidential features, we turn to exploiting them for predicting question response time. As described in Section III, we consider two prediction tasks. Task 1 aims to classify questions based on whether they will receive an answer within the median response time, i.e. 16 minutes, or not. On the other hand, Task 2 tries to separate questions that are answered within an hour from those that take more than one day to receive an answer.

Since we split the data based on median response time, the class sizes are exactly the same for Task 1 (∼1.7 million questions/class). For Task 2, the class sizes are skewed, with most questions (63.5%) having response time less than 1 hour. If we simply predict all questions having low response time, we could achieve low prediction error, although such a trivial method is not very informative. To address this issue, we follow the under-sampling technique [6] so as to balance the class sizes (∼344 thousand questions/class). We perform the sampling several times and report averaged results.

We perform each task based on various settings depending on the set of features used in prediction. Specifically, we experiment with (i) all but only non-tag based (10) features, (ii) (1) tag-based feature tag_popularity, (iii) another (1) tag-based feature percent_subs_t, (iv) all but only tag-based (9) features, and finally (v) all (19) features. We also employ several different classifiers to eliminate the pitfall of interpreting results based on only a specific classifier type. In particular, we use two linear classifiers: logistic regression (Log Reg) and SVM with linear kernel (SVM (Lin)), as well as two non-linear classifiers: decision tree (DT) and SVM with radial basis function kernel (SVM (RBF)).

In Fig. 9 we show the classification performances (based on 10-fold cross validation) on Task 1, measured by both accuracy and F1 score (bars depict standard deviation across folds). The results are quite intriguing. First and foremost, non-tag based features perform inferior to all the other settings. Each of the two models learned based on single tag-based features outperforms the models learned using all 10 non-tag features by a significant margin.

Second, we observe that percent_subs_t proves to be more evidential than tag_popularity. This is expected, since the former feature is based on the answering speed of responsive "subscribers" of tags. Using all of the 9 tag-based features improves accuracy by 7-9% and the F1 score by 4-5% over using only percent_subs_t. Adding all non-tag features on top of all tag-based features, however, incurs only a slight improvement, by only another 1-2%. This difference is insignificant specifically for the non-linear models.

Finally, we note that these observations remain qualitatively the same across different types of classifiers, which indicates that the utility of our proposed tag based features compared to the non-tag based features is not an artifact of the choice of a specific classifier.

The same arguments hold for results on Task 2, as shown in Fig. 10. Here, the performance is slightly better than Task 1, as the separation between response times of the class instances is larger and hence the task is relatively easier.

To further analyze the importance of tag based features, we quantify the discriminative power of all the features in estimating response time. In particular, we use their sum of information gains weighted by the number of samples split by each feature at the internal tree nodes [7] based on the decision tree models. We present the top 10 features for Task 1 and Task 2 in Table III ranked by their importance. We observe that the most discriminative (top-3) as well as the majority (6/10) of features are the ones based on tags.

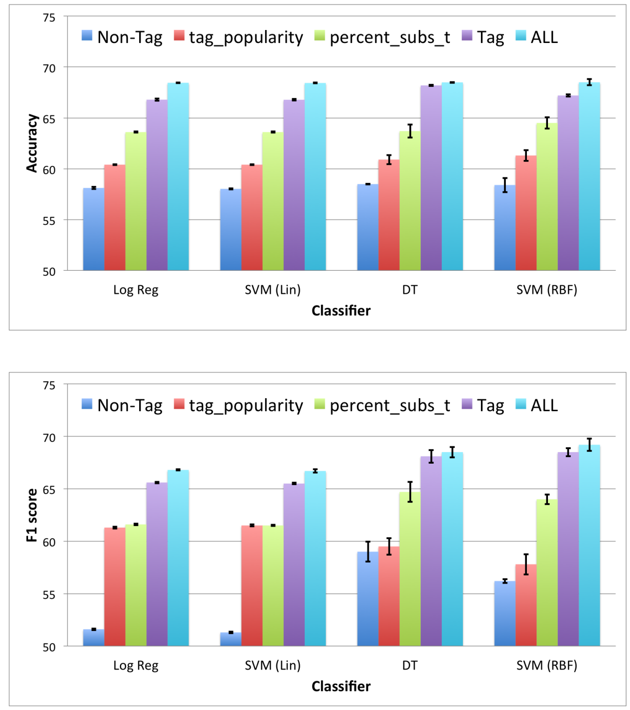


Fig. 9. Classification performance based on (top) accuracy and (bottom) F1 score on Task 1. All non-tag based features yield lower performance than tag popularity feature alone. Tag based features provide the largest boost in prediction performance.

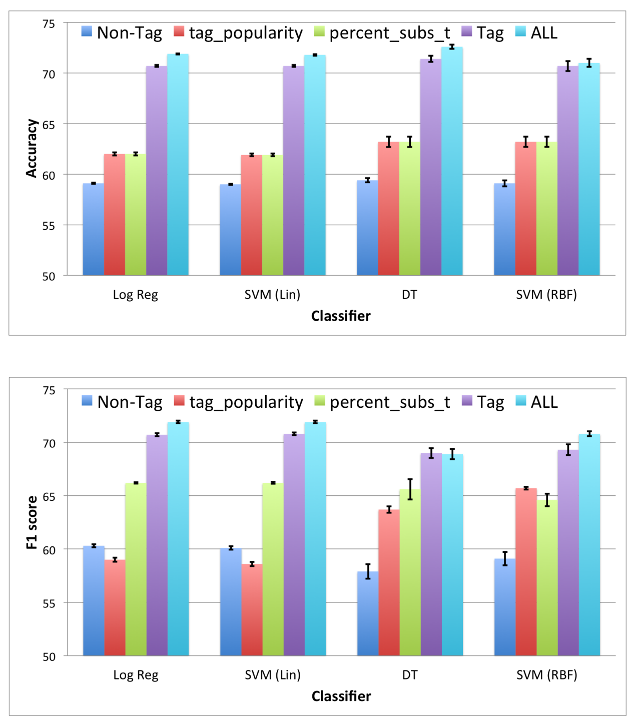


Fig. 10. Classification performance based on (top) accuracy and (bottom) F1 score on Task 2. All non-tag based features yield lower accuracy than tag popularity feature alone. Tag based features provide the largest boost in prediction performance.

TABLE III
TOP 10 MOST IMPORTANT FEATURES BASED ON INFORMATION GAIN. ASTERIKS (*) INDICATES TAG BASED FEATURES. NOTICE THAT THE MAJORITY OF THE FEATURES AS WELL AS TOP 3 MOST PREDICTIVE FEATURES ARE TAG BASED.

Task 1		**Task 2**	
* percent_subs_t	0.440	* percent_subs_t	0.506
* tag_popularity	0.173	* percent_subs_ans	0.266
* num_subs_t	0.130	* tag_popularity	0.085
body_len	0.123	body_len	0.057
* percent_subs_ans	0.033	* num_subs_ans	0.030
* num_subs_ans	0.026	end_que_mark	0.013
* tag_specificity	0.025	title_len	0.010
end_que_mark	0.013	* num_subs_t	0.010
title_len	0.013	code_len	0.009
code_len	0.012	* tag_specificity	0.007

Finally, we note that the prediction performance is slightly better for non-linear models for both tasks. The highest accuracy (F1 score) on Task 1 is 68.5% (69.2%) and 72.6% (71.9%) on Task 2. While these predictions are significantly better than the random baseline, performance is not tremendously high. Nevertheless, our study clearly demonstrates the predictive power of tag based features, which could be utilized in future efforts of estimating response time.

VI. RELATED WORK

Asaduzzaman *et al.* [8] explained why questions remain unanswered on Q&A sites and proposed a categorization of questions based on features such as too short, too hard, program specific, fails to attract expert, etc. Their analysis does not identify tags as one of the factors that determine if a question would remain unanswered. Mahmud *et al.* [2] studied how the response time can be predicted for questions asked on Twitter. Since Twitter is not specifically meant to be a Q&A site, their results do not take into consideration the information that is available on such sites.

Rechavi *et al.* [1] analyzed average response time and other facts in Yahoo! Answers and predominantly looked for what constitutes a best answer on the site. They found that in most cases the first received answer is marked as the best answer by the asker, implying the askers looking for quick responses, while the community carefully chooses the best answer amongst all the answers received. The authors also found that being a follower or fan of a person on Yahoo! Answers does not yield quicker response from the person. Avrahami *et al.* [3] analyzed the response time in instant messaging. They built various features from the messaging app and desktop settings, and predicted if a response will be given to a message within a certain period. In addition, Sun *et al.* [9] developed link prediction methods not only to infer whether a relational link will form but also when the link is expected to be formed in the future, although their work is on network data and not on question response time.

Wang *et al.* [10] studied Quora to understand the impact of its site design and organization on the growth and quality of its knowledge base. Different from earlier behavioral studies, this work focuses on the design aspects of the Q&A sites and their effects on user engagement. Other related works on Q&A sites include the quality and value analysis of questions and answers. Harper *et al.* [11] studied the predictors of answer quality with respect to two dimensions; site characteristics (e.g., type and organization of communities and experts), and question characteristics (e.g., strategies like thanking in advance and showing prior effort). Anderson *et al.* [12] analyzed the factors that contribute to the long-term value of questions on StackOverflow.

Another group of works studies the lifespan of users in online Q&A sites. Yang *et al.* [13] analyzed three large Q&A sites from three countries to understand the predictive patterns in participation lifespans of users. Arguello *et al.* [14] studied user communities to understand the contributing factors to success in their ability to respond to and retain active participants. Several other works also looked at newcomers' retention [15], [16], [17]. Finally, Movshovitz *et al.* studied StackOverflow to build an understanding of its reputation system and user contributions [18]. None of these earlier works studied the response time of questions however.

VII. CONCLUSION

Question response times on Q&A sites affect the lifespan of these online communities as faster responses increase user satisfaction and engagement. We study a large dataset from StackOverflow to identify evidential factors associated with response time. We conjecture that the tags chosen by the askers influence how their questions are answered. Our in-depth data analysis and prediction experiments demonstrate the competence of tag-based features as well as their superiority over more obvious, previously studied non-tag based factors.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Army Research Office Young Investigator Program grant under Contract No. W911NF-14-1-0029, an R&D grant from Northrop Grumman Aerospace Systems, an Office of Naval Research SBIR grant under Contract No. N00014-14-P-1155, and Stony Brook University Office of Vice President for Research. Any findings and conclusions expressed in this material are those of the author(s) and should not be interpreted as representing the official policies or views, either expressed or implied, of the funding parties.

REFERENCES

[1] A. Rechavi and S. Rafaeli, "Not all is gold that glitters: Response time & satisfaction rates in Yahoo! answers." in *SocialCom/PASSAT*. IEEE, 2011, pp. 904–909.

[2] J. Mahmud, J. Chen, and J. Nichols, "When will you answer this? estimating response time in twitter." in *ICWSM*. The AAAI Press, 2013.

[3] D. Avrahami and S. E. Hudson, "Responsiveness in instant messaging: predictive models supporting inter-personal communication." in *CHI*. ACM, 2006, pp. 731–740.

[4] M. R. Morris, J. Teevan, and K. Panovich, "What do people ask their social networks, and why?: a survey study of status message Q&A behavior," in *SIGCHI*, 2010.

[5] M. E. J. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[6] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 875–886.

[7] P. G. Neville, "Decision Trees for Predictive Modeling," *SAS Institute Inc.*, 1999.

[8] M. Asaduzzaman, A. Mashiyat, C. Roy, and K. Schneider, "Answering questions about unanswered questions of stack overflow," in *10th Working Conference on Mining Software Repositories. Mining Challenge*, 2013.

[9] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: Relationship prediction in heterogeneous information networks." in *WSDM*. ACM, 2012, pp. 663–672.

[10] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: an analysis of Quora," in *WWW*, 2013.

[11] F. M. Harper, D. R. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online Q&A sites." in *CHI*. ACM, 2008, pp. 865–874.

[12] A. Anderson, D. P. Huttenlocher, J. M. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow." in *KDD*, 2012, pp. 850–858.

[13] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic, "Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities." in *ICWSM*, 2010.

[14] J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, , and X. Wang, "Talk to me: foundations for successful individual-group interactions in online communities," in *CHI*, 2006.

[15] C. Lampe and E. W. Johnston, "Follow the (slash) dot: effects of feedback on new members in an online community." in *GROUP*, 2005, pp. 11–20.

[16] E. Joyce and R. E. Kraut, "Predicting continued participation in newsgroups." *J. Computer-Mediated Communication*, vol. 11, no. 3, pp. 723–747, 2006.

[17] G. Dror, D. Pelleg, O. Rokhlenko, and I. Szpektor, "Churn prediction in new users of yahoo! answers." in *WWW*, 2012, pp. 829–834.

[18] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow." in *ASONAM*, 2013, pp. 886–893.