Analysis of Factors Influencing User Contribution and Predicting Involvement of Users on Stack Overflow

Maliha Mahbub
University of Saskatchewan
Saskatoon, Canada
mam789@mail.usask.ca

Najia manjur
University of Saskatchewan
Saskatoon, Canada
nam907@mail.usask.ca

Mahjabin Alam
University of Saskatchewan
Saskatoon, Canada
natasha.mahjabin@gmail.com

Julita Vassileva
University of Saskatchewan
Saskatoon, Canada
jiv@cs.usask.ca

ABSTRACT
Active involvement of new community members is essential for Q&A platforms such as Stack Overflow, to make the platform efficient and more inclusive. However, more than half of Stack Overflow users contribute only once and disappear. This decreases the diversity of viewpoints and experience on the platform. This paper aims to identify factors that can discourage users from active participation after their first or second post. We collected a dataset of the responses to questions posted by new users (answers, comments, upvotes, downvotes) and analysed the tone of the feedback and its impact on the users' ongoing participation. We considered as new users those who registered to Stack Overflow for the last two years before the data collection and classified then into three groups based on the number of their posts (low, medium and high number of posts) on Stack Overflow. The differences in the responses between the three user groups have been validated by performing one-way ANOVA and Pearson's chi-square test. Based on these results we trained a machine learning model using a SVM classifier which predicts whether a user is likely to post or not with an accuracy of 88.69 %. Our work contributes to identifying and quantifying the potential underlying factors behind the decline in participation and dropout of new users on Stack Overflow.

Keywords
Stack Overflow, User Analysis, One-way ANOVA, SVM Classifier, User Prediction Model

1. INTRODUCTION
Stack Overflow (SO) is one of the most popular Q & A based platforms for programmers, having over 50 million monthly visitors[7], over 16 million questions[19] and 19 million answers[11]. SO has detailed guidelines for posting questions and some fundamental standards for providing feedback to the questions. Expert users who have been using Stack Overflow for a long time are rewarded with badges and reputation [14]. In order to garner a high reputation on the site, the user must be active on a regular basis on the SO platform and their questions must have many positive responses. However, novice users may not have the correct vocabulary or expertise to formulate a technical question. As a result, they may end up getting negative responses such as "stupid question" or "This is such common issue, just google". Such negative responses to posts may discourage users to limit or seize their contributions to the platform. This kind of users make up for almost half the users of the platform [18]. Unsolved questions in SO have seen an exponential growth over the years [16] and they continue to be an issue for new users who seek help. Studies have shown that online trolling and negative responses worsen over the active time of a user in a community [8] and 77% of users tend to ask for help on the SO platform only once [13]. In this paper, we identify and validate the factors which impact the state of participation of the new users on Stack Overflow. "New users" are those who have been registered to SO for less than 2 years until August 2019 (when the dataset was created). We selected and analysed five features from the SO post responses to understand how getting little to no response and negative responses is related to users posting behaviour. Based on our findings, we built a machine learning classifier to predict posting status of users in SO.

2. RELATED WORK
The Stack Overflow or SO platform has turned into a valuable resource for both skilled and amateur programmers for glitches, bug and any code related problems. Managing such a large user base has been a challenge and an ongoing topic for investigation and research from different perspectives. Anderson et al. [3] explores the correlation between user reputation and quality of answers and its impact on the design of the site. Asaduzzaman et al. [4] mines the unanswered questions in SO to reveal the underlying factors that lead to questions remaining unanswered, such as title length, askers' score, post length etc. Alharthi et al. [2] investigates sev-

eral factors that impact the quality of questions in SO and predicts the score of the question, which indicates its overall quality. Similar idea of prediction has been explored in Shao et al. [17] developed a prediction model which analyses the latent context of a question and recommends an answer for the user. Calefato et al. [6] developed a framework based on successful questions on SO to provide an evidence -based guideline for programmers to write better questions in SO. Grant et al. [12] explores the use of badge, to motivate users. When a question has better wording and quality, it attracts more users and the user gets upvotes which in turn helps the score. Adaji et al. [1] investigates specific social support strategies that influence users to contribute in SO More recent studies have focused on the behavioural and personality traits of SO users as well, in order to target the emotional aspects of the users who ask questions [15, 5]. The novice or infrequent users who just started out face some level of criticism or neglect by the more experienced users on SO, a phenomenon related to maintaining community boundaries by hazing. Hazing is a psycho-social phenomenon where the newcomers in a tightly knit group face backlash and elitist attitude (which is sometimes borderline abusive)[9]. Slag et al. [18] discusses the difficulties encountered by "one day flies", users who post only once in their profile's lifetime and do not contribute to the platform afterwards. Our work further investigates the effect of the factors identified in [18] by providing statistical and empirical validation to the hypothesis proposed in [18].

3. RESEARCH QUESTIONS AND DATASET

Since we wanted to compare the responses of the posts, not the nature of post itself, we eliminated two of Slag et al. [18] factors: duplicate questions and uncommon tags. Instead, to further investigate the features of questions asked by such inactive users, we added five new factors: the number of upvotes on a question (Up Votes), the downvotes (Down Votes), the number of comments on a post (Comment Count), the reputation of users (Reputation), and the types of comments on a post (Comment Texts). We chose to add reputation since it affects how a user's posts is perceived by other users. We aim to answer two research questions:
"1. Do these factors have any quantifiable relation to the frequency of posts of users in Stack Overflow?"
"2. Can we predict whether a user will drop out and stop posting?".

3.1 Data Collection

We collected data from Stack Exchange Data Explorer [1], an open source tool to collect publicly available data from Stack Overflow. We used Stack Exchange Data Explorer to collect information about users who created their profile on Stack Overflow in 2017.

For our work, we chose to consider only the questions posted by users as their contribution. We collected the number of answers, comments, upvotes, downvotes, view count given against (received by) each post of a user and the user's reputation. We decided to analyze the mean values of these features for each user so that we can consider all of them in a normalized form since the distribution of responses is not equal for all users. In order to determine the overall tone of a comment, we inferred the polarity of the comments from the text of the comments through sentiment analysis on the text by using TextBlob on all the comments received by a user. Polarity generally falls within the range of -1 to 1 where -1 refers to a very negative sentiment, 0 refers to a neutral sentiment and 1 refers to a very positive sentiment. We categorized this range into five different groups as shown in Table 1.

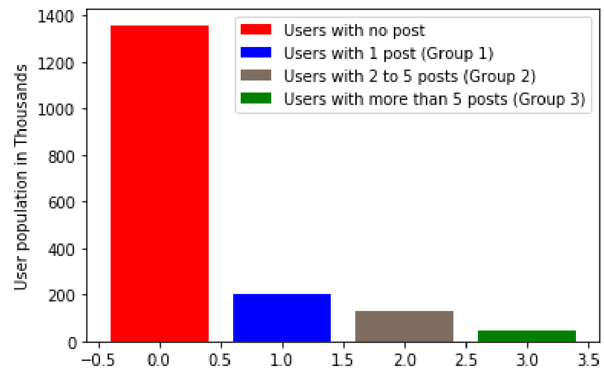

Figure 1: Distribution of users population (in thousands) based on the number of posts they made on Stack Overflow

Table 1: Group based on overall polarity of the comment

Group	Description
1 (Highly negative)	If -1 <= polarity <-0.5
2 (Moderately negative)	If -0.5 <= polarity <0
3 (Neutral)	0
4 (Moderately positive)	If 0 <polarity <= 0.5
5 (Highly positive)	If 0.5 <polarity <= 1

3.2 Target user groups

The users were categorized into the following groups based on the number of questions they posted from January 2017 to June 2019: group 1 (users who posted a question once), group 2 (users who have posted from 2 to 5 times), group 3 (users who have posted more than 5 times). The distribution of user population according to the number of posts they made in their lifetime on SO in out collected data is depicted in the figure 1. We aim to identify such users by analyzing their past experiences on SO. In our prediction model, described in Section 5, for predicting the future behavior of a user, group 1 is labelled as the negative class whereas groups 2 and 3 are combined into a single category as the positive class. Therefore, the labeled classes considered for this study are:

1. Negative class: Users who will discontinue making any contribution to Stack Overflow after their first post.

2. Positive class: Users who will continue contributing to the platform.

4. DATA ANALYSIS

[1]https://data.stackexchange.com/stackoverflow/query/new

	View Count	AnswerCount	CommentCount	Reputation	DownVote	UpVote	Polarity
View Count	1	0.32	0.41	0.1	0.3	0.51	0.24
AnswerCount	0.32	1	**0.86**	0.1	**0.59**	**0.68**	**0.67**
CommentCount	0.41	**0.86**	1	0.11	**0.57**	**0.61**	**0.79**
Reputation	0.1	0.1	0.11	1	0.08	0.24	0.1
DownVote	0.3	**0.59**	**0.57**	0.08	1	**0.52**	0.53
UpVote	**0.51**	**0.68**	**0.61**	0.24	**0.52**	1	0.49
Polarity	0.24	**0.67**	**0.79**	0.1	**0.53**	0.49	1

Figure 2: Distribution of users population (in thousands) based on the number of posts they made on Stack Overflow

The data analysis of feature selection, validation of metrics and prediction model is provided below.

4.1 Feature Selection

From the data collected from SO, we performed Pearson correlation analysis to find out how strongly each feature is related to another. The results of correlation analysis for the features are shown in figure 2. Following the Cohen's classification system [10], only the largest relationships i.e. where the correlation coefficient r>0.5, have been considered to be significantly correlated. From figure 2, it is evident that the correlation coefficient of comment count, answer count, downvote, upvote, and polarity are significant i.e. greater than 0.5. Therefore, for our feature analysis, these five features are selected as the final metrics for next stage.

4.2 Feature Distribution in User Groups

To answer the research question: Do these features have any quantifiable relation to the frequency of posts of users in Stack Overflow?, we analyzed their statistical differences among the three user groups. We used a one-way ANOVA test and Pearson's chi-square test to establish the statistical evidence of the differences in terms of the features among the three user groups.

All of the five features are plotted against the number of posts from users. And the plotting is done for each of the three target user groups to observe the difference of plots in each groups. The sections below provide in-depth description of each of the features on all three user groups.

4.2.1 Number of Answers against Number of Posts

Figure 3 depicts the distribution of average number of answers against the number of posts from each user from the target group of users on SO. The mean number of answers among three groups are: 1.10 (group 1), 3.31 (group 2) and 15.70 (group 3), which indicates that users in group 3 receive significantly more responses to their posts compared to users in groups 1 and 2. The p-value in one-way ANOVA test indicates significant statistical difference among the three groups in terms of the mean number of answers they receive against their posts (F(2,375196) = 678.8, p = .000).

4.2.2 Number of Comments against Number of Posts

The mean number of comments among three groups are: 2.22 (group 1), 6.60 (group 2) and 30.00 (group 3), which indicates that the mean number of comments significantly increases with the increasing number of posts in each group. Moreover, it is also evident from figure 4 that users who posted less (no more than five times) exhibited a higher tendency of receiving no comments from other users. The result of one-way ANOVA test indicates significant statistical difference between the groups than within the groups in terms of number of comments they receive against their posts (F(2,375196) = 187.3, p = .000).

4.2.3 Number of Upvotes against Number of Posts

The graphs show the relation between the average number of upvotes with the number of posts in Figure 5. The mean upvotes in group 1 and 2 are significantly lower than group 3 (0.696, 1.853 and 8.877 respectively), which indicates that the posts made by the users of group 3 are more appreciated and receive higher number of upvotes than the posts made by the users who are less active. One-way ANOVA result indicates significant statistical difference among the three groups in terms of number of upvotes they receive against their posts (F(2,375196) = 17.15, p = .000)

4.2.4 Number of Downvotes against Number of Posts

From figure 6, it can be observed that the mean number of downvotes in group 1 and group 2 (0.548 and 1.309 respectively) are lower than that of group 3 (4.17). This means that the users who are posting more questions are also getting fewer downvotes. This is an important and surprising observation since a higher number of posts could have also led to increased number of downvotes, which turns out to not be the case. One-way ANOVA result indicates significant statistical difference between the groups than within the groups in terms of number of comments they receive against their posts (F(2,375196) = 473.04, p = .000).

4.2.5 Comment Polarity against Number of Posts

Table 2: Percentage of each comment polarity category received by the user groups

Polarity Group	User Group		
	1	2	3
1 (Highly negative)	0.2%	0.1%	0%
2 (Moderately negative)	20.1%	20.7%	13.6%
3 (Neutral)	24.4%	11.7%	1.3%
4 (Moderately positive)	48.4%	67.2%	85.1%
5 (Highly positive)	1.1%	0.3%	0%

The result of cross tabulation in Table 2 revealed that users from group 3 received zero highly negative comments, and the least proportion of moderately negative comments. They also received the highest proportion of moderately positive comments (85.1%). On the other hand, the users of group 1 received the lowest amount of moderately positive comments(49.5%) among the user groups. It can be concluded from Table 2 that with the increase in the number of posts made by the users, there is an increase in the positive comments and decline in the negative remarks received by the post owners. Lastly, the result of Pearson's chi-square test establishes the statistically significant relationship between the polarity of comments and the user groups ($\chi^2(8, 297447) = 20310.29, p = 0.000$).

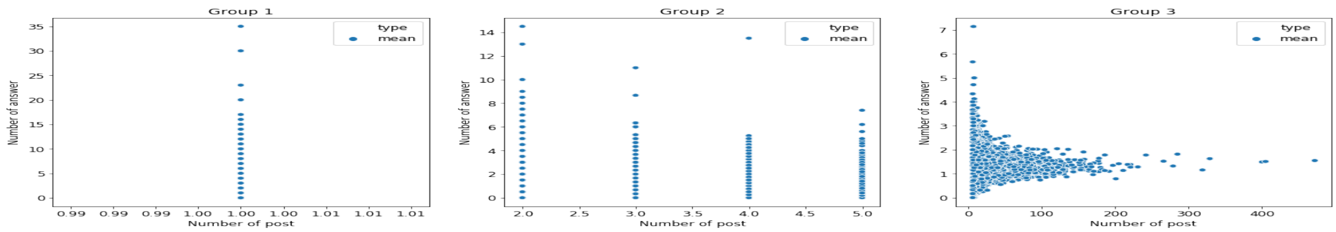

Figure 3: Distribution of average number of answers against number of posts among three user groups

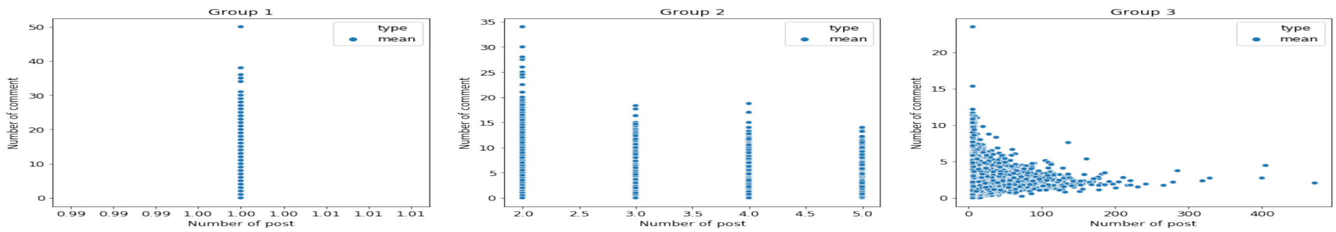

Figure 4: Distribution of average number of comments against number of posts among three user groups

	Name	Measure
Features	Mean answer	Scale
	Mean comment	Scale
	Mean upvote	Scale
	Mean downvote	Scale
	Mean polarity	Scale
Target	User Class	Nominal

Table 3: Attributes of the dataset employed in the SVM classifier

Classifier	Accuracy	Precision	Recall
SVM	0.8869	0.9875	0.7635

Table 4: Prediction model performance per evaluation metric

5. PREDICTING USER PARTICIPATION

Based on the findings from correlation analysis and statistical testing of factors influencing users post frequency, we developed an actual prediction model to answer the second research question. In order to develop and train our model, we took advantage of a popular supervised machine learning algorithm called support vector machines (SVM). Table 3 describes the features and target groups employed in our model. We divided the original data set into training set representing 80% of the data and testing set representing the remaining data. By using the five features, we divide our users into two classes: likely to post and not likely to post. The performance of our prediction model i.e. how well it predicts the user class is evaluated using the metrics: accuracy, precision and recall. Our model performs significantly well and yields a high score in terms of all three metrics as illustrated by Table 4.

6. IMPLICATIONS AND FUTURE WORK

From our data analysis, we observed that a low number of answers and comments, a high number of downvotes and negative comments, and a low number of upvotes are more prevalent in the posts of users who have posted fewer times compared to the users who have higher number of posts. Since these users are receiving negative remarks and downvotes even with fewer posts, this may play a role in discouraging them from seeking help again from SO. In the light of this discovery, we trained a SVM classifier model with the five special features and divided the users into two classes: users who will post in the future and users who will not. The model has shown good performance with high accuracy and effectiveness.

Previous works which mostly focused on how users can ask better questions or build a better profile to attract more answers to their questions. The novelty of our study is to identify infrequent users and find a possible factor underlying their withdrawal, so that the community owner/ moderator can make the platform more welcoming and less hostile for them. Our study has some limitations as well. We could not consider the number of deleted questions of a user as one of the factors that could contribute to users' decline in posts since Stack Exchange Data Explorer does not provide that data. The research also lacks a qualitative analysis from feedback of infrequent or absent users. Therefore, as part of our our future plan, we will attempt to explore the user modelling of infrequent posting through a targeted qualitative user study of SO users.

7. CONCLUSIONS

More than half of the users in Stack Overflow tend to ask for help on the platform only once and never post again. In this paper, we identified five main features / metrics which we hypothesized to be related to the inactive status of users. We collected the responses to posts in SO for users who have their SO profiles for 2 years (2017 to 2019) and selected five factors with strong correlation. Ours statistical analysis supports our hypotheses and validates the effect of these factors having a significant correspondence to users' posting frequency. Using these factors as selected features, we trained a machine learning model that predicts whether or not a user will post in the Stack Overflow platform, based

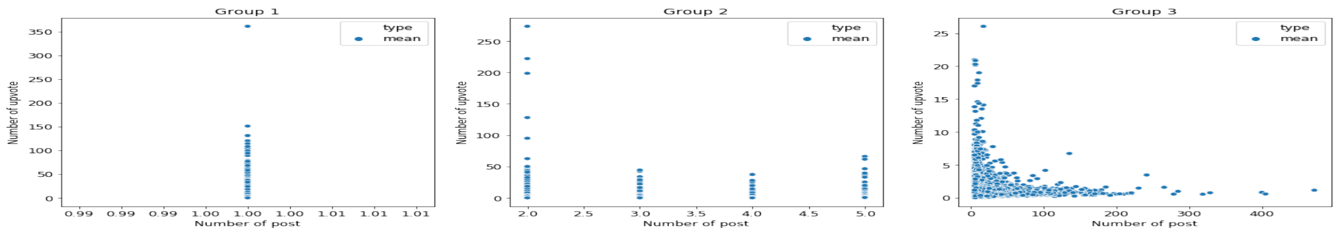

Figure 5: Distribution of average number of upvotes against number of posts among three user groups

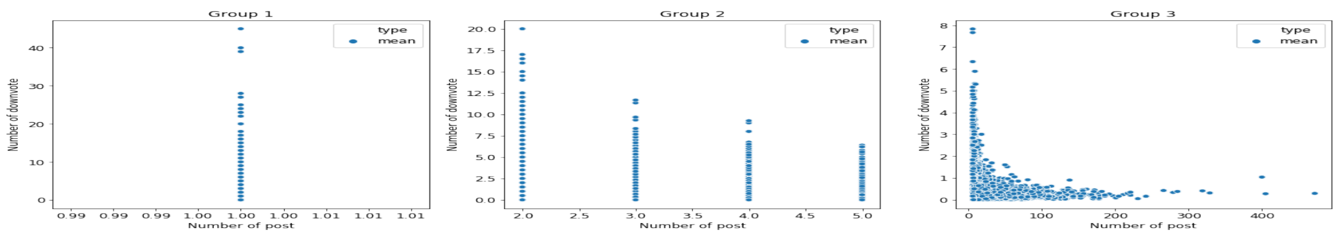

Figure 6: Distribution of average number of downvotes against number of posts among three user groups

on the responses their posts received till now. This prediction can identify users who have reduced their posting in SO and face lack of encouragement and thus can benefit from a positive nudge, help or mentorship. The significance of the contribution of our analysis and prediction model is that it can help to provide more equitable treatment of newcomers, and thus increase the diversity of the SO community.

8. REFERENCES

[1] I. Adaji and J. Vassileva. Towards understanding users' motivation in a q&a social network using social influence and the moderation by culture. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 349–350, 2017.

[2] H. Alharthi, D. Outioua, and O. Baysal. Predicting questions' scores on stack overflow. In *2016 IEEE/ACM 3rd International Workshop on CrowdSourcing in Software Engineering (CSI-SE)*, pages 1–7. IEEE, 2016.

[3] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858, 2012.

[4] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider. Answering questions about unanswered questions of stack overflow. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 97–100. IEEE, 2013.

[5] B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of stackoverflow users. In *2013 IEEE international conference on software maintenance*, pages 460–463. IEEE, 2013.

[6] F. Calefato, F. Lanubile, and N. Novielli. How to ask for technical help? evidence-based guidelines for writing questions on stack overflow. *Information and Software Technology*, 94:186–207, 2018.

[7] P. Chatterjee, M. Kong, and L. Pollock. Finding help with programming errors: An exploratory study of novice software engineers' focus in stack overflow posts. *Journal of Systems and Software*, 159:110454, 2020.

[8] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[9] A. Cimino. The evolution of hazing: Motivational mechanisms and the abuse of newcomers. *Journal of Cognition and Culture*, 11(3-4):241–267, 2011.

[10] J. Cohen. Statistical power analysis. *Current directions in psychological science*, 1(3):98–101, 1992.

[11] Craig Smith. Interesting stack overflow statistics and facts (2020). https://expandedramblings.com/index.php/stack-overflow-statistics-and-facts/, 2020. [Online; accessed 25-March-2020].

[12] S. Grant and B. Betts. Encouraging user behaviour with achievements: an empirical study. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 65–68. IEEE, 2013.

[13] John Slegers. The decline of stack overflow. https://hackernoon.com/the-decline-of-stack-overflow-7cb69faa575d, 2015. [Online; accessed 26-March-2020].

[14] 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 *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 886–893. IEEE, 2013.

[15] N. Novielli, F. Calefato, and F. Lanubile. Towards discovering the role of emotions in stack overflow. In *Proceedings of the 6th international workshop on social software engineering*, pages 33–36, 2014.

[16] M. M. Rahman and C. K. Roy. An insight into the unresolved questions at stack overflow. In *2015*

IEEE/ACM 12th Working Conference on Mining Software Repositories, pages 426–429, May 2015.
[17] B. Shao and J. Yan. Recommending answerers for stack overflow with lda model. In *Proceedings of the 12th Chinese Conference on Computer Supported Cooperative Work and Social Computing*, pages 80–86, 2017.
[18] R. Slag, M. de Waard, and A. Bacchelli. One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 458–461. IEEE, 2015.
[19] Wikipedia contributors. Stack overflow — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Stack$_Overflowoldid = 940723804, 2020.[Online; accessed 20 - February - 2020]$.