An Empirical Study on Neophytes of Stack Overflow: How Welcoming the Community is towards Them

Abdullah Al Jobair, Suzad Mohammad, Zahin Raidah Maisha, Md. Jubair Ibna Mostafa and Md. Nazmul Haque

Department of Computer Science & Engineering,
Islamic University of Technology, Boardbazar, Gazipur, Bangladesh

Keywords: Neophyte, Stack Overflow (SO), New User, Hostile Environment.

Abstract: Stack Overflow (SO) is the most popular question and answers (Q&A) platform for programmers with a rapidly expanding community of new users. However, the unwelcoming environment towards new users has been under discussion for several years, which is a major concern towards the enhancement of a skillful community. In this work, we study a specific group of users who are either registered in the last 45 days or have a reputation less than or equal to 50 and term them as *"neophytes"*. We investigate whether neophytes actually face hurdles while collaborating in Stack Overflow and, if so, identify the reasons behind this phenomenon by qualitative and quantitative analysis. Our study finds that neophytes are indeed facing hurdles while collaborating in the platform. The reasons behind the hurdles include harsh moderation of posts, negligence of the posts, deleting or closing of posts, downvoting without providing any proper reasoning, etc. Our findings can provide guidelines to create a more user-friendly Stack Overflow community. Furthermore, this study can guide researchers to observe the reactions of neophytes in adverse situations and recommend some steps for the community to make positive changes to the Stack Overflow environment.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The exponential growth of the software development industry leads to forming a community for aiding one another with wisdom and experience. Q&A Platforms are the result of such a need which eventually establishes a community for sharing knowledge. In the community, users share skills and techniques among themselves to solve different problems. Among all the online software development Q&A platforms, Stack Overflow is the largest and the most renowned one (May et al., 2019). From the dawn of its origin, a total of 16.5 million users have registered on the site with an average of 3,370 new users registering every day and making around 11,203 posts on a daily basis[1] (based on a query run in August 2021). Today's massive repository of 21 million questions and 31 million answers in Stack Overflow (Moutidis and Williams, 2021) is the result of the gradual progress of the community since 2008.

The accessibility of this extensive dataset has brought about a number of researches on this platform[2] including evolution of community, posts, code snippets, along with user behavior, user participation, mining SO and associated technologies and many more (Ahmed and Srivastava, 2017; Adaji and Vassileva, 2016). However, a limited number of studies are focused on the environment of the community for new users.

With the community's swift expansion (Mamykina et al., 2011), Stack Overflow's environment draws significant attention. Any hostile nature of the community may turn off the eagerness for participation, which hinders the lively ambience of the platform. However, related studies intimate the existence of unwelcoming environments specially to the new users. Less experienced users become frustrated due to the obscurity in closing questions (Tóth et al., 2020). It ultimately leads the community to become hostile and unsupportive, mostly to the new users. The study of Abbas presented unanswered questions, negative feedbacks and deleted questions as the root of a massive discouraging impact towards users (Abbas, 2019). According to (Slag et al., 2015), 47% of users

[1] https://data.stackexchange.com/stackoverflow/query/1541382

[2] https://stackoverflow.blog/2009/06/04/stack-overflow-creative-commons-data-dump/

Al Jobair, A., Mohammad, S., Maisha, Z., Mostafa, M. and Haque, M.
An Empirical Study on Neophytes of Stack Overflow: How Welcoming the Community is towards Them.
DOI: 10.5220/0011081100003176
In *Proceedings of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2022)*, pages 197-208
ISBN: 978-989-758-568-5; ISSN: 2184-4895

Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

post only once and disappear from the community. They found that new users' posts get removed more often in addition to not receiving responses to their questions at a higher rate. Along with researchers, the SO community itself is concerned about this pressing issue[3]. The yearly site satisfaction survey of the community[3] presents the unwelcoming environment as the top frustrating and unappealing factor for SO users. The following quote from the survey result reflects this situation of the community,

> *"The toxic nature of the community .... Scares people from even signing up let alone asking questions"*

In this study, we address this issue by validating whether the unwelcoming nature of the SO is a reality and if so, investigate the probable reasons for encountering such a hostile environment by the new users. To achieve our goal, these new users have been distinguished from the total users and termed as *"neophytes"*. To validate our conjectures and properly understand the neophytes' state in SO, this study addresses the following research questions-

1. **RQ-1: Do Neophytes Face Hurdles while Collaborating in Stack Overflow?**
   The allegation of the Stack Overflow environment being unwelcoming and hostile, specially to the neophytes, is a persisting problem for the community. Our aim is to verify whether the problem exists in reality or not. The affirmative outcome of this research question led us to investigate the second research question.
2. **RQ-2: What are the Potential Reasons for Neophytes Facing Hurdles while Collaborating in Stack Overflow?**
   There could be several potential reasons for which neophytes are facing hurdles. Identifying those reasons will help to understand the unwelcoming nature of the platform and provide insight towards solving the problem.

Our study validates the problem of the unwelcoming environment of SO, specially to the neophytes. In addition, we also find a number of reasons, including posts being deleted, closed, posting duplicated questions or answers, community rules violation etc. for facing a hostile environment.

The rest of the paper is structured as follows. Section 2 discusses the motivation behind our work. Section 3 defines the term "neophyte" and describe their characteristics. Section 4 discusses the related works on SO, specifically new users of SO. Section 5 presents the methodology of this study, including data extraction, qualitative and quantitative analysis procedures. Section 6 represents the results and analysis, answering RQ-1 and RQ-2 and thereafter recommending some steps for Stack Overflow community. Section 7 discusses the validity of the study and section 8 concludes the paper by outlining future work.

2 MOTIVATION

The unwelcoming community has been a buzzing issue since the very beginning. The ambiguous closure of posts (Tóth et al., 2020), negative feedbacks (Cheng et al., 2014), offensive language (Cheriyan et al., 2020) make the platform hostile, specially affecting the new users. The gradual posts of renowned blog sites, official surveys, posts of Meta Stack Exchange[4] and official blogs of Stack Overflow community itself vocalizes the continual nature of the issue. A blog post of *the exception catcher*[5] claims Stack Overflow as a difficult community for participation by observing the frequent downvoting tendency to a post.

Meta Stack Exchange is a Q&A site where users discuss the workings of SO. Here topic of each question falls under some specific tags. The most upvoted post[6] on *"new-user"* tag urges the community to be supportive to new users. It has been viewed 45 thousand times and received 1,830 upvotes (according to August 2021) which clearly represents that the community is not welcoming enough for new users.

Jay Hanlon, the former Executive Vice-President (EVP) of culture and experience of SO is also vocal about the issue and asks for the prompt change of the situation[7].

> *"Stack Overflow isn't very welcoming. It's time for that to change."*

Furthermore, a qualitative accumulation of evidence by Slegers[8] provides a verdict on the hatred nature of SO. The major point of this accumulation is *"Stack Overflow hates new users"*.

The hostile nature of SO is not a problem of recent days, rather the situation has been prevailing since long ago and no improvement is reflected according to the developer survey of the community of 2019[9] and

[3] https://stackoverflow.blog/2020/01/22/the-loop-2-understanding-site-satisfaction-summer-2019/

[4] https://meta.stackexchange.com/

[5] https://theexceptioncatcher.com/blog/2012/09/stackoverflow-is-a-difficult-community-to-participate-in/

[6] https://meta.stackexchange.com/questions/9953/

[7] https://tinyurl.com/424h7w4j

[8] https://hackernoon.com/the-decline-of-stack-overflow-7cb69faa575d

[9] https://insights.stackoverflow.com/survey/2019

2020[10]. The survey of 2019 in Figure 1, expresses that there is no progress in the welcoming environment of the community because 73% developer votes as the environment remained the same as it was in last year, 2018. Whereas, the survey of 2020 in Figure 2 shows 70.6% vote.

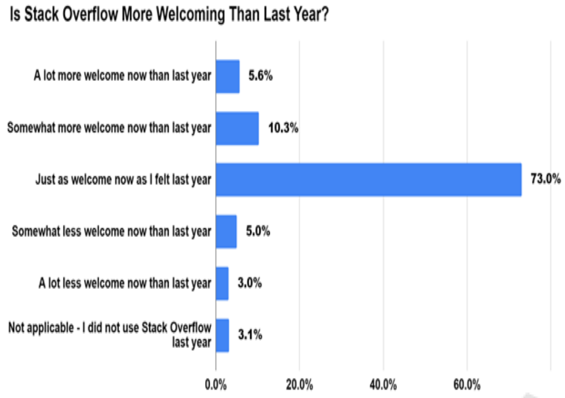

Figure 1: Developer Survey 2019.

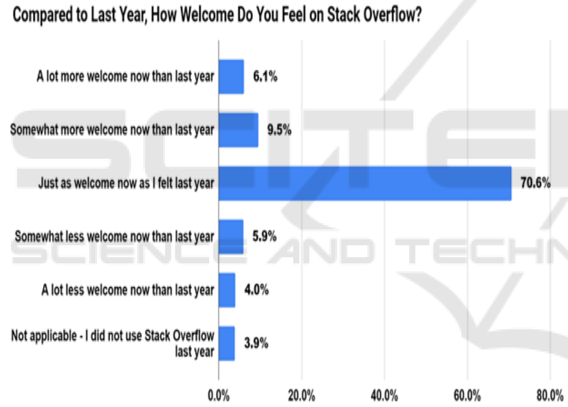

Figure 2: Developer Survey 2020.

All the studies, surveys, blogs and meta discussions substantiate the claim of a hostile environment, specially for the new users. Moreover, the issue has been persistent for years and sufficient research works are not documented on this problem which encouraged us to work on this concern.

3 DEFINING NEOPHYTE

The contributions of various levels of users, starting from the professionals to the novices, make Stack Overflow so lively, dynamic and the most used question-answer site (May et al., 2019). To evaluate how new users contribute to the community and to analyze how welcoming the environment is towards them, our research is concentrated on a fixed group of users who are termed as *"Neophytes"*.

According to our definition, neophytes are those groups of users who are either registered in Stack Overflow in the last 45 days or have a reputation of less than or equal to 50. Users registered in SO within the last 45 days are taken into consideration because Stack Overflow defines them as *"New users"*[11]. Although it is an acceptable indication of newly joined users to the platform, it does not specify anything on their contribution to the site. Thus, a reputation constraint is integrated to inspect neophytes' contribution to the platform.

After rigorous analysis, we end up with two reputation boundaries of 38 and 50 reputation.(Slag et al., 2015) worked with 38 reputation as they found it the average reputation of medium active users. But a user with 38 reputation lacks the privilege of commenting which is a vital feature[12]. On the contrary, Stack Overflow allows almost all basic operations like questioning, answering, commenting, upvoting (apart from downvoting which is assigned for the reputed users) if someone gets to 50 reputation[11]. So, to ensure an impactful presence of users in SO, 50 reputation is chosen over 38.

Therefore, if one of the conditions gets satisfied for a user, that user will be considered as a neophyte. Everyone other than neophytes in SO is specified as *"regular users"* throughout the paper.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Algorithm 1: Algorithm to find neophytes from registered user pool.

```
1: procedure FindingNeophytes($reg\_users$)
2:     $neophytes = []$
3:     for each $user$ in $reg\_users$ do
4:         if ( then($user.reputation \leq 50$) or ($user.registration\_day \leq 45$) )
5:             $neophytes$.add($user$)
6:         end if
7:     end for
8:     return $neophytes$
9: end procedure
```

Algorithm 1 dissociates neophytes from all the registered users pool in Stack Overflow. It receives "*reg_users*" as a parameter which represents registered users of SO. The output is the list of "*neophytes*" separated from the registered users. In line-2, an empty list of neophytes is taken. For each user in registered users, the constraints of 50 repu-

[10] https://insights.stackoverflow.com/survey/2020

[11] https://meta.stackexchange.com/questions/310881/

[12] https://stackoverflow.com/help/privileges

tation boundary or the registration date within last 45 days is checked in line-4. One fulfilling the constraints is added to the neophytes list.

According to our definition, 89.9%[13] (14,897,718) of total users of Stack Overflow data dump 2020 are neophytes. For this study, we focus on a specific group of users to investigate the attitude of SO community towards them.

4 RELATED WORK

Many research works are conducted after the SO dataset is made public. The research works of Stack Overflow are diverse. The works are done on a regular basis on numerous domains.

"Post" analysis is one of the richest domains with research studies from the very beginning of Stack Overflow. According to (Abric et al., 2019), duplicate posts are mostly posted by inexperienced users. With the advancement of technology, the number of obsolete answers are increasing. This issue was addressed in a study by (Zhang et al., 2021a) and the authors observed that some tags are prone to obsolete answers. The community is also indifferent to specific tags that keep those tag-specific questions unanswered (Saha et al., 2013). A study on new users by (Hart and Sarma, 2014) denied the claim of new users relying on intrinsic factors (answerer's reputation, representation of answer etc.) only to identify answer quality. Reasons of closing a post can be categorized into 5 groups (Tóth et al., 2020). However, the ambiguous closure of post is a concerning issue which frustrates and hurts the users specially the newbies and make the environment feel hostile to them. These studies include vital information like impact of closure of questions and new users' perspective of detecting quality. But the works lack anything related to how new users' posts are accepted to the community.

Analysis of *"Comment"* is another vital aspect to understand the environment and culture of a platform. Studies like categorizing the comments (Sengupta and Haythornthwaite, 2020) indicate how the comments help in learning and increasing skills. One of the recent studies on SO investigated how the platform manages comments (Zhang et al., 2021b). Analyzing these comments can provide with insights on gender hospitality in Stack Overflow (Brooke, 2019). A study on norm violations in SO shows that its comments are offensive and unwelcoming by presenting a taxonomy of norms that are violated (Cheriyan et al., 2020). Unfortunately, the domain still requires research on comments, specially addressing the situation of new users.

Various studies on user badge, reputation, participation have been making the SO *"User"* domain enriched since the very dawn of its establishment. A study by (Yanovsky et al., 2021) discussed the association of user contribution and behavior with achievement of badges. A much needed contribution for the new users is the research of (Bosu et al., 2013), where they provided guidance to new users on enhancing their reputation swiftly. An associative study on involvement of habits of individuals with high and low reputation (Movshovitz-Attias et al., 2013) presents that extremely high-reputed users are the dominant source of replies, particularly high-quality responses. On the contrary, low-reputed users ask a bulk of questions on the site without answering any question. In another study, the reputation and contribution of a user against the completeness of their profile had been investigated (Adaji and Vassileva, 2016). The study observed that users with complete profiles have relatively high reputation and also they post high quality contents.

A number of studies present the concern on the environment of Stack Overflow. A research on detecting and classifying offensive language claims SO as unwelcoming by using offensive language (Cheriyan et al., 2021). In an earlier study, the authors investigated a group of users labelled as *"one-day fly"* which refers to users who never returned after posting only once (Slag et al., 2015). They examined why one-day flies don't contribute to the site more than once. In spite of discarding the allegation that new users - (i) post frequent duplicate questions, (ii) post on uncommon tags and (iii) get less views, they found new users posts frequently get removed and remain responseless. A subsequent study on one-day flies (Abbas, 2019) discusses some elements, which contribute to the major issue of inactive users in SO. The author employs a comprehensive literature review strategy to develop the analysis. An investigation on *"Slashdot"* (a news and discussion site) finds that it has established a distributed moderating mechanism to offer input on the merit of its posts (Lampe and Johnston, 2005). This research looks at three different theories for how new users learn to join in a digital community: learning transfer from past experiences, observation of other members, and feedback from other members. Another investigation on four big comment-based news communities depict that negative feedback causes major behavioral changes that are harmful to the community (Cheng et al., 2014).

While all the studies mentioned above contribute

[13] https://data.stackexchange.com/stackoverflow/query/1384160/

to significant aspects of user reputation, badge, participation and community environment in Stack Overflow, to the best of author knowledge there is limited study dedicated towards the environment neophytes are facing and how they feel for the platform. In our empirical study we want to investigate the environment of the platform for neophytes. It will be a stepping stone towards building a friendly skillful community and enhancing the quality of the huge knowledge base, Stack Overflow is aiming for.

5 METHODOLOGY

An overview of our research methodology is presented in Figure 3. At first, we extract the neophytes' and regular users' data from the official SO data dump and then qualitative as well as quantitative analyses are performed to answer the RQ-1 and RQ-2. The qualitative analysis is performed considering posts, blogs and surveys along with the manual analysis. The qualitative analysis is followed by the quantitative analysis. For quantitative analysis, a query based statistical analysis is conducted. The following subsections describe our method with further details.

Data Extraction

For the analysis, the June 2020 database of Stack Overflow was initially used, provided by Stack Exchange Archive[14]. But we had to choose an alternate data source as some tables crucial for our analysis were missing from the data dump (e.g. CloseAsOffTopicReasonTypes, CloseReasonTypes, FlagTypes, PendingFlags, SuggestedEdit, PostFeedback etc.). These tables are concrete indicators of the difficulty neophytes face in the community, without which the study loses a significant amount of gravity.

So, as an alternative, *"Stack Exchange Data Explorer"* has been used. Stack Exchange Data Explorer is an online platform that facilitates the execution of SQL queries on public data from the Stack Exchange Network[15]. We accessed the data in August 2021 at the time of our analysis. An overview of our dataset is presented in Table 1.

In order to understand the status quo, a timeframe of 1st January 2020 to 31st December 2020 is selected for the queries. As the data of 2021 was not complete during the analysis, the 2021 dump was omitted to avoid anomalies. The one-year timeframe provided a total of 4,456,062 posts. Out of this total post, 1,161,701 posts are posted by 619,171 neophytes. This expresses a neophyte posted on average 1.88 posts. Whereas 3,294,361 posts are posted by 458,745 regular users which increases the ratio to 7.18 posts per regular user. As most of the related works used data spanning from 6 months to 1 year, it can be said that the data used is sufficient. Python is used for our statistical analysis. Our entire analysis has been done using SQL Server as Stack Overflow uses SQL Server databases to store their data dumps.

Table 1: An overview of dataset.

Dataset	
Total Posts	4,456,062
Neophytes' Posts	1,161,701
Data Duration	1st January - 31st December, 2020

These data are used to answer our research questions that are previously mentioned in the "Introduction" section. In the following parts, we distinctly explain the analyses and approaches to answer the research questions.

RQ-1: Do Neophytes Face Hurdles while Collaborating in Stack Overflow?

The very first step to our analysis is confirmation of the conjecture that Neophytes are facing hurdles while collaborating in SO. This issue concerns a significant portion of SO users rather than just a few of them. Both qualitative and quantitative analysis is performed to answer this research question.

Qualitative Analysis

A qualitative analysis has been conducted to examine whether neophytes actually face hurdles in Stack Overflow. Meta Stack Exchange Posts on *"new user"* tag, Blogs and surveys[7,3] by Stack Overflow community, renowned blog sites[5,8] contain the evidences and insights on unwelcoming environment of Stack Overflow. As part of the investigation, we analyze these documents to understand some factors that influence creating the hostile environment. In addition, a manual analysis has also been conducted on neophytes' posts to strengthen our qualitative study. The outcome of manual analysis gives us an initial impression regarding this problem and validates the research process which has been presented in *"Result of RQ-1"*. Our intuition, influenced by the resources mentioned above, indicates the existence of a real problem between regular users and the neophytes.

In order to prevent biases, 300 neophytes who

[14] https://archive.org/download/stackexchange

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

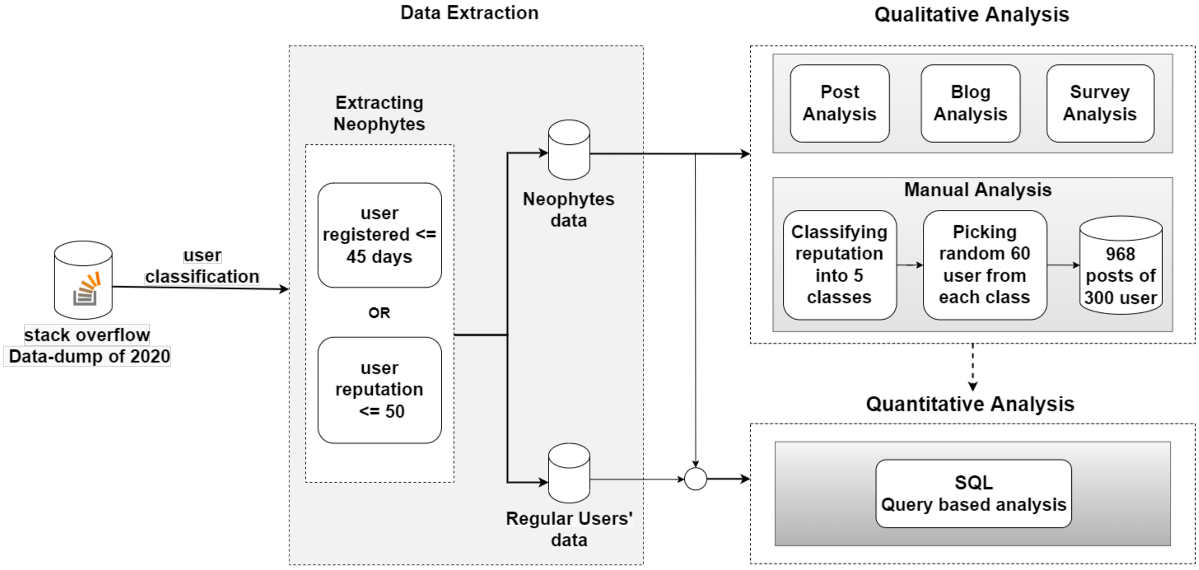

Figure 3: An overview of methodology.

have registered in 2020 are randomly selected for our manual analysis. The analysis has been performed on their 968 posts. The intention is to find out how frequently neophytes face unwelcoming situations while collaborating to inspect the claim of their hurdles.

At the beginning, the reputation boundary of neophytes i.e. 50 points is clustered into 5 classes considering the upper and lower bounds, each class has a difference of 10 reputations. (0-10, 11-20, 21-30, 31-40, 41-50 reputation). Then we have randomly picked out 60 users from each class in order to avoid any biases which resulted in a set of 300 neophytes. The manual analysis result provides convincing stats on neophytes facing hurdles in SO which is presented in the result section of RQ-1.

Quantitative Analysis

For validating the research question (RQ-1), a query-based quantitative analysis is performed on *"Stack Exchange Data Explorer"*. We have analyzed a comparative investigation on the total number of neophytes with total users and neophytes' posts in contrast to total posts. Moreover, how many posts of neophytes are getting downvoted and whether they face hurdles in their first posts while their first posts are the downvoted ones are also investigated.

RQ-2: What are the Potential Reasons for Neophytes Facing Hurdles while Collaborating in Stack Overflow?

To answer this research question, a qualitative analysis is performed which is followed by a quantitative analysis. The qualitative analysis discovers some probable reasons whereas the quantitative analysis validates those reasons.

Qualitative Analysis

To investigate the reasons, first a qualitative analysis is performed. The same data of 968 posts of 300 neophytes that we have accumulated are also analyzed here. This time along with further analysis of the posts, the neophytes profiles have been considered as well to answer RQ-2. The profile based analysis leads us to understand the gradual activities of them and help recognize the reasons for their hurdles. We investigate the total number of posts by a neophyte, the date difference of their first and last post and the date difference of the most down-voted post and the immediate next post. The goal is to inspect the activity of the neophytes after facing some unwelcoming situations. We also investigate their badges and overall progress, mentioning our observations on their posts as well as their profiles. From this qualitative analysis, several potential reasons are identified for neophytes facing hurdles in Stack Overflow.

Quantitative Analysis

To delicately investigate and verify the obtained list of reasons, a query-based quantitative analysis is conducted. As previously mentioned in the *"Data Extraction"* section, the Stack Overflow data dump of 2020 is used in this quantitative analysis. We have formulated numerous query based questions and executed the queries in the online query site of Stack Overflow, the Stack Exchange Data Explorer. The outcomes of the query significantly confirmed our manual analysis.

6 RESULT ANALYSIS

The outcome of the questions will give us a clear pictorial view of the issue. Thus, will lead to the community thinking about taking effective and long term steps towards resolution.

Result of RQ-1

According to the qualitative analysis, among 968 posts of neophytes, 254 posts are negatively scored among which 123 posts have no explanation or proper cause of getting the negative score. In addition, 47 posts are duplicate, 64 are being closed and surprisingly 110 posts get no response at all. These 110 posts have got no comments, no response along with 0 score count. The analysis outcome is precisely depicted in Figure 4.

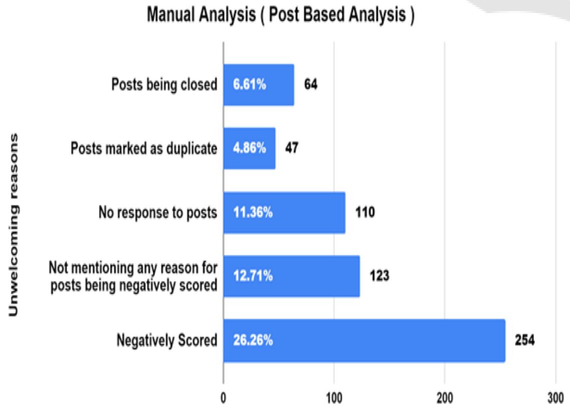

Figure 4: Post based Manual analysis.

The total number of neophytes' posts that are facing difficulties (posts being closed, posts marked as duplicate, no response to posts, negative scored posts) is about 49% of the total analyzed posts. As 'posts with no explanation or proper cause of getting the negative score' i.e. 123 posts are already included in 'negative scored posts' i.e. 254 posts, we do not include it to total 49%. Almost half of our randomly selected neophytes posts denote that they are being neglected, resulting in the community being unwelcoming towards them.

According to our quantitative analysis, neophytes posted 26.07% of the total posts of the platform in 2020 which is a significant portion of the total community. Out of these 26.07% posts i.e. 1,161,701 posts a total of 108,568 posts get negative score which is 9.35% of the total posts of neophytes. The percentage might be misleading as it is small and seems very normal. But the same ratio for the regular users comes down to 3% only. It clearly depicts the difference of posts getting negatively scored for neophytes and regular users.

The aforementioned statistics that we have gathered from qualitative analysis lead us to the conclusion that neophytes are facing hurdles in SO.

Key Findings of RQ-1:
According to the qualitative analysis, about 49% of the accumulated 968 posts of neophytes are found to be having difficulties like being duplicated, closed, negatively scored etc. Quantitative analysis reports a 9.35% of the total posts of neophytes getting negatively scored which is only 3% for regular users. The analysis vividly represents that neophytes are facing hurdles while collaborating in Stack Overflow.

Result of RQ-2

Among manual analysis of 300 neophytes' profiles depicted in Figure 5, total 77 neophytes obtained the *"Informed"* badge which means only 25.67% undergo the entire tour page to gain knowledge on how Stack Overflow works.

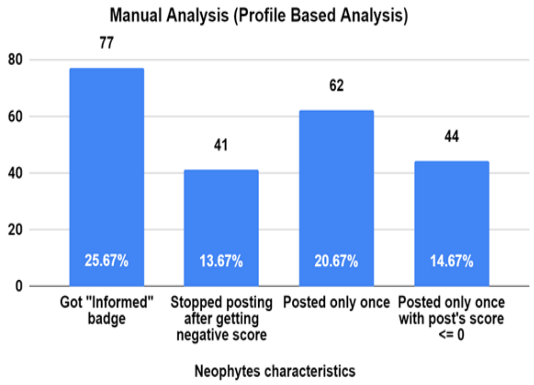

Figure 5: Profile based manual analysis.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

A total of 41 neophytes did not post further after their posts were negatively scored which results in 13.67% of our accumulated data-set. Moreover, 62 neophytes posted only once, defined as *"one-day-flies"* (Slag et al., 2015), which is 20.67% of the total 300 neophytes. Among those 62 neophytes, 44 of them got a score less than or equal to 0 in their posts.

From the qualitative analysis, 9 potential reasons have been identified. Each of these reasons is responsible for neophytes facing hurdles in Stack Overflow. The reasons for neophytes facing hurdles in the community are -

- Posts being closed
- Posts marked as duplicate
- Not mentioning any reason for posts being negatively scored
- No response to posts
- Unaware of Stack Overflow rules and culture
- Deletion of posts
- Moderation without proper reasoning
- Rude comments
- Steep learning curve

From our query based quantitative analysis on Stack Overflow data dump 2020, the statistics vividly depicts the presence of the reasons. Among these introduced reasons, several reasons (Posts being closed, Posts marked as duplicate, No response to posts, Not mentioning any reason for posts being negatively scored) have been validated by the quantitative study.

As previously mentioned in the subsection *"RQ-2: What are the potential reasons for neophytes facing hurdles while collaborating in Stack Overflow?"* of ***"Methodology"***, a number of queries are formed and executed for the quantitative analysis of this study. The queries of Listing 1 and Listing 2 are two of those queries.

```
select count (p.Id)
from Posts p
inner join Users u
on p.OwnerUserId=u.Id
inner join PostLinks pl
on pl.PostId=p.Id
where (u.Reputation<=50 or u.CreationDate>=getdate()-45)
and (p.CreationDate between datefromparts(2020,01,01) and datefromparts(2020,12,31))
and (pl.LinkTypeId=3)
```

Listing 1: Query to find posts marked as duplicate.

```
select count(p.Id)
from Posts p
inner join Users u
on u.Id=p.OwnerUserId
left outer join PendingFlags pf
on pf.PostId=p.Id
left outer join SuggestedEdits se
on se.PostId=p.Id
where (u.Reputation<=50 or u.CreationDate>=getdate()-45)
and (p.CreationDate between datefromparts(2020,01,01) and datefromparts(2020,12,31))
and (p.CommentCount=0 and p.Score<0 and p.ClosedDate is null and pf.PostId is null and se.PostId is null)
```

Listing 2: Query to find posts with no reason for being negatively scored.

The SQL in Listing 1 represents a query to find the number of duplicate posts of neophytes. If a post has the property *linkTypeId* equal to 3, then this post is counted as a duplicate post which has a prior post very similar to it. The SQL in Listing 2 represents query to find the number of posts having no explanation or proper cause of getting the negative score. A post with negative score having neither any comment, pending flag, suggested edit nor they are being closed is considered to be negatively scored without reasoning. As every closed post has a reason associated with it, so the closed post constraint is also integrated into the query.

However, some reasons (Deletion of posts, Moderation without proper reasoning, Rude comments, Steep learning curve) could not be analyzed by our query due to lack of necessary data. Stack Overflow does not make these data publicly available. The inaccessibility of all these data imposes a barrier to validate them quantitatively.

Table 2: Comparison of total posts and neophytes posts.

Unwelcoming Reasons	Total Posts	Neophytes Posts
Posts being closed	104,461	52,761 (50.5%)
Posts marked as duplicate	78,652	38,508 (48.96%)
Negative scored posts	207,508	108,568 (52.32%)
Not mentioning any reason for posts being negatively scored	56,717	25,421 (44.8%)
Posts got no response at all	892,557	212,457 (18.29%)

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 2 shows a comparison between total posts vs neophytes' posts. Whereas, Figure 6 depicts the stats of several reasons from the quantitative analysis.

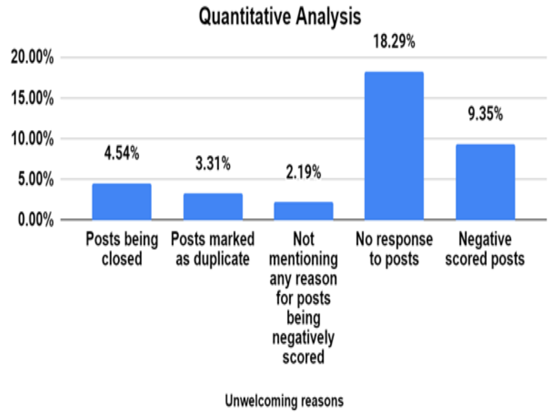

Figure 6: Quantitative analysis.

Posts being Closed

From queries conducted on the 2020 dump, we have observed that total 104,461 posts were closed in 2020. Out of these closed posts, a significant portion of 50.5% (52,761 posts) belong to neophytes which are half of the total closed posts, as presented in Table 2. Figure 6 shows that 4.54% of total posts of neophytes are getting closed. On the contrary, it is only 1.57% for regular users' posts. Although the percentage in comparison with their huge post count seems usual, there is a clear difference between the ratio of neophytes and regular users. Such actions affect neophytes as a result lose their enthusiasm and interest from further contributing to the site.

Posts Marked as Duplicate

Our analysis in Table 2 states that, in 2020 total 78,652 posts were marked as duplicate where 38,508 (48.96%) posts belong to neophytes. According to Figure 6, 3.31% of the total posts of neophytes are marked as duplicate. The percentage declines to 1.3% for the regular users.

Duplicate posts generally receive negative feedback from the community. However, (Abric et al., 2019), in their research, depicts that duplicate questions and answers contain some unique information that benefits the asker. Even if it is marked as a duplicate question, the original question does not serve the purpose of the asker. This causes frustration for the neophyte because they did not get help as well as faced harsh moderation on top of it.

Not Mentioning Any Reason for Posts being Negatively Scored

From Table 2, we can see that in total, 56,717 posts of 2020 data dump got negatively scored but no reason (comment, suggested edit, flag) was there to show-cause the down-vote. Out of which 25,412 posts were posted by neophytes which are 44.8% of these 56,717 posts. Compared with the total number of posts(1,161,701 posts) of neophytes in 2020, the amount is 2.19% as per depicted in Figure 6. Although the amount seems to be small, such behavior strongly demotivates neophytes from further contributing to the site. Downvoting posts is definitely one of the mechanisms that helps in maintaining the quality of the platform. But if it is done without explanation of what went wrong with the post, it fails to serve the purpose.

Posts Got No Response at All

18.29% (212,457 posts out of total 1,161,701 posts of neophytes in 2020) of neophytes remained completely responseless which is presented in Figure 6 and Table 2. The posts are neither being closed nor received any answer. Even those posts do not contain any comment, edit suggestion or any flag. Amidst the 209,025 unique neophytes whose post got no response, 112,486 neophytes (53.81%) did not post further. The alarming percentage hints at how this culture affects the neophytes.

Unaware of Stack Overflow Rules and Culture

Neophytes often make irrelevant answers, security vulnerable solutions, opinion-based questions, ask for debugging and violate Stack Overflow rules. All these are because of being unaware of SO rules and culture. Neophytes are often not familiar with the conventions in Stack Overflow which leads to miscommunication between neophytes and regular users. A significant number of 2,174,619 neophytes (15.15%) do not go through the SO tour page and ultimately lack the *"Informed"* badge. From the regular users' perspective, this hampers the integrity of SO as the site gets overflowed with repetitive and unnecessary posts. However, the response from this dynamic often discourages neophytes from engaging in any further discussions.

Deletion of Posts

(Slag et al., 2015), one day fly's posts account for 15.4% of overall post deletions. The study also discusses how the post deletion system can contribute to lessened participation of one-day-flies. (Abbas, 2019) discussed *"Deleted Questions"* as one of the significant factors for people not participating in SO.

As Stack Overflow keeps all the information related to deletion of posts private[16], it is quite impossible to make any quantitative analysis on deleted posts. However, an idea can be generated regarding the deletion of posts by counting the number of neophytes getting the *"Peer Pressure"* badge. The *"Peer Pressure"* badge is obtained when users delete their own post with a score of -3 or lower. The quantitative analysis informs a total of 153,515 neophytes having *"Peer Pressure"* badge in 2020.

Moderation without Proper Reasoning

In SO, users get responses within a very short period of time, typically within 21 minutes[17]. Moderation in Stack Overflow is so fast that their questions face negative responses, closure or deletions etc. within a very short period of time, like in less than ten minutes[17]. This can easily lead to users getting frustrated. Thus it is one of the vital factors which makes communication between regular users and neophytes difficult.

Rude Comments

Rude Comments are flagged and deleted quickly, but even in that situation, users end up reading the rude comments against them. This makes neophytes who are not yet accustomed to the culture of Stack Overflow, feel frustrated and unwelcoming. During our analysis of individual users' profiles, we found several cases that indicate that a neophyte has stopped posting after they received negative responses to their posts. Rude comments towards neophytes dissatisfy them leading them to leave the community.

Steep Learning Curve

Stack Overflow is different from most question and answer platforms as they aim to create an effective knowledge base of developers. To maintain such effectiveness, participating in SO requires a high learning curve. That leads to the point that understanding the purpose of SO or participating properly in the community takes time. By that time, neophytes are flooded with downvotes, closure deletion and many other forms of negative response.

Key Findings of RQ-2:
Potential reasons for neophytes facing hurdles in Stack Overflow are - posts being closed, posts marked as duplicate, not mentioning any reason for posts being negatively scored, no response to posts, misconception of Stack Overflow rules and culture, deletion of posts, moderation without proper reasoning, rude comments, steep learning curve.

Recommendation

With the qualitative and quantitative analysis, it is evident that proper collaboration and initiatives are necessary from both neophytes' and Stack Overflow's ends to better the environment of SO. We recommend some steps for the Stack Overflow community.

- For closed posts, SO is recommended to use a pre-post automated prediction tool that can predict whether a post will be closed or not before the post is published. It will also predict the reasons as well as respective suggestions for closing posts and notify the user. As a result, users can realize their flaws in posts and act according to the suggestions. By this, the number of closed posts will also lessen in SO.
- For posts being downvoted without mentioning any reason, SO should impose the moderators and privileged users to mention proper reasons for downvotes. The reasons for such moderation can help users identify and rectify their flaws.
- The rude comments need to be detected before they are published publicly. That is, comments should be verified through SO moderation before posting. The moderation can be proactively performed by an automated tool. This will conceal any rude language from the sight of users and ultimately will reduce the level of hostility.
- For posts that got no response at all, SO should take steps to detect post quality and encourage privileged users to review them. Moreover, an automated tool can be designed to route the post to the more suitable users. The reviews will be notified to the owners so that these can positively guide them.
- Neophytes should be more cautious about their posts. They should follow the rules and regulations of Stack Overflow as well as accustom themselves to the culture. There should be an assessment that ensures that the users are acquainted with the rules.

[16] https://stackoverflow.com/questions/56770820/

[17] https://meta.stackexchange.com/questions/61301/

7 THREATS TO VALIDITY

Internal Validity

- Stack Overflow does not disclose any data regarding the deleted posts. The only way to obtain this information is to import earlier data and compare it to the present one, which is not a valid concrete work as well. Due to the absence of this data, our research lacked a quantitative investigation on this reason.
- Stack Overflow does not provide any data on vital information like closed posts, flags and suggested edits of posts in the Stack Exchange Archive(offline database). This led us to work with the online version of SO data dump (Stack Exchange Data Explorer). Due to the rapid update of online data-dumps, we have to perform the analysis binding a particular time-frame constraint to avoid the possible anomalies in our data.

External Validity

- In order to maintain consistency we limited our study on Stack Overflow only. So, the research outcome may not reflect the condition of other Q&A sites like reddit, quora etc. An analysis on these sites is also required to understand the overall condition of new users and the environment for them.
- Only the database of 2020 has been considered for our analysis to understand the environment neophytes face in Stack Overflow. A database of pre-pandemic period (before 2019) could be compared with a database of pandemic period. It will indicate if there is any effect of covid pandemic on the neophytes characteristics and environment of Stack Overflow.

8 FUTURE WORK & CONCLUSION

Unwelcoming behavior towards neophytes has been under discussion for many years, with little steps taken related to it. The study sheds light on this issue by confirming its validity and identifying significant reasons behind this problem by providing definitive data and statistics. The findings will help build a welcoming environment by uniting all ranges of users. It will encourage new users to be actively involved in this knowledge base.

A user with a reputation within 50 is considered as neophytes. A further clustering of this group based on their activeness would give detailed insights about their characteristics. This will guide in future works to see the distinction among active and inactive users. In addition, it will indicate the ratio of neophytes having hurdles while participating in the platform.

Sentiment analysis on neophytes would be an effective study, along with understanding the impact of comments on neophytes' posts. The extensive studies will lead to the most appropriate suggestions for Stack Overflow to resolve this problem.

REFERENCES

Abbas, A. E. (2019). Investigating 'one-day flies' users in the stackoverflow: Why do and don't people participate? In *2019 International Conference on ICT for Smart Society (ICISS)*, volume 7, pages 1–5.

Abric, D., Clark, O. E., Caminiti, M., Gallaba, K., and McIntosh, S. (2019). Can duplicate questions on stack overflow benefit the software development community? In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 230–234.

Adaji, I. and Vassileva, J. (2016). Towards understanding user participation in stack overflow using profile data. In *International Conference on Social Informatics*, pages 3–13. Springer.

Ahmed, T. and Srivastava, A. (2017). Understanding and evaluating the behavior of technical users. a study of developer interaction at stackoverflow. *Human-centric Computing and Information Sciences*, 7(1):1–18.

Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., and Kraft, N. A. (2013). Building reputation in stackoverflow: An empirical investigation. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 89–92.

Brooke, S. (2019). "condescending, rude, assholes": Framing gender and hostility on stack overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2014). How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Cheriyan, J., Savarimuthu, B. T. R., and Cranefield, S. (2020). Norm violation in online communities–a study of stack overflow comments. *arXiv preprint arXiv:2004.05589*.

Cheriyan, J., Savarimuthu, B. T. R., and Cranefield, S. (2021). Towards offensive language detection and reduction in four software engineering communities. In *Evaluation and Assessment in Software Engineering*, pages 254–259.

Hart, K. and Sarma, A. (2014). Perceptions of answer quality in an online technical question and answer forum.

SCIENCE AND TECHNOLOGY PUBLICATIONS

In *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*, CHASE 2014, page 103–106, New York, NY, USA. Association for Computing Machinery.

Lampe, C. and Johnston, E. (2005). Follow the (slash) dot: Effects of feedback on new members in an online community. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, page 11–20, New York, NY, USA. Association for Computing Machinery.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. (2011). Design lessons from the fastest q&amp;a site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2857–2866, New York, NY, USA. Association for Computing Machinery.

May, A., Wachs, J., and Hannák, A. (2019). Gender differences in participation and reward on stack overflow. *Empirical Software Engineering*, 24(4):1997–2019.

Moutidis, I. and Williams, H. T. (2021). Community evolution on stack overflow. *Plos one*, 16(6):e0253010.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). 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.

Saha, R. K., Saha, A. K., and Perry, D. E. (2013). Toward understanding the causes of unanswered questions in software information sites: A case study of stack overflow. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2013, page 663–666, New York, NY, USA. Association for Computing Machinery.

Sengupta, S. and Haythornthwaite, C. (2020). Learning with comments: An analysis of comments and community on stack overflow. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Slag, R., de Waard, M., and Bacchelli, A. (2015). 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.

Tóth, L., Nagy, B., Gyimóthy, T., and Vidács, L. (2020). Why will my question be closed? nlp-based presubmission predictions of question closing reasons on stack overflow. In *2020 IEEE/ACM 42nd International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 45–48.

Yanovsky, S., Hoernle, N., Lev, O., and Gal, K. (2021). One size does not fit all: A study of badge behavior in stack overflow. *Journal of the Association for Information Science and Technology*, 72(3):331–345.

Zhang, H., Wang, S., Chen, T.-H., Zou, Y., and Hassan, A. E. (2021a). An empirical study of obsolete answers on stack overflow. *IEEE Transactions on Software Engineering*, 47(4):850–862.

Zhang, H., Wang, S., Chen, T.-H. P., and Hassan, A. E. (2021b). Are comments on stack overflow well organized for easy retrieval by developers? *ACM Trans. Softw. Eng. Methodol.*, 30(2).

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS