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

Authors

Suzad Mohammad (170042024)
Abdullah Al Jobair (170042030)
Zahin Raidah Maisha (170042032)

Supervisors

Md. Jubair Ibna Mostafa
Lecturer
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Md. Nazmul Haque
Lecturer
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

A thesis submitted to the Department of CSE in partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering

Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh.
May, 2022

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by under the supervision of Md. Jubair Ibna Mostafa, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh and Md. Nazmul Haque, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Suzad 10.05.22

Suzad Mohammad
Student ID: 170042024
Academic Year: 2017-2018
Date: 10 May, 2022

Abdullah Al Jobair 10-5-22

Abdullah Al Jobair
Student ID: 170042030
Academic Year: 2017-2018
Date: 10 May, 2022

Zahin Raidah Maisha 10/05/22

Zahin Raidah Maisha
Student ID: 170042032
Academic Year: 2017-2018
Date: 10 May, 2022

Approved By:

Supervisors:

10.05.2022

Md. Jubair Ibna Mostafa
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

10.05.2022

Md. Nazmul Haque
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

Dedicated to our parents

Table of Contents

List of Figures

List of Tables

Acknowledgment

All praise and gratitude be to Allah, the most beneficent, the most merciful. Without His mercy and blessings it would not been possible.

We would like to express our grateful appreciation for **Md. Jubair Ibna Mostafa**, Lecturer, Department of Computer Science & Engineering, IUT and **Md. Nazmul Haque**, Lecturer, Department of Computer Science & Engineering, IUT for being our adviser and mentor. Their motivation, suggestions and insights for this research have been invaluable. With their support and proper guidance this research become possible. Their valuable opinion, time and input provided throughout the thesis work helped us to do our thesis work in proper way. We are really grateful to them.

We are also grateful to **Mohayeminul Islam**, Assistant Professor, Department of Computer Science & Engineering, IUT and **Tajkia Rahman Toma**, Assistant Professor, Department of Computer Science & Engineering, IUT for their valuable inspection, suggestions and guidance.

Lastly, we thank every member of Department of Computer Science & Engineering for their immense support and cooperation throughout the work. Their helping hands have made what we have achieved today possible.

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 and hindrance 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"*. Upon establishing significance of the definition of neophytes, we perform manual analysis of neophytes' posts. We organize our research work into two research questions where we investigate whether neophytes actually face hurdles while collaborating in Stack Overflow and, if so, identify the potential 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 SO 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.

Acceptance of Work

Conference

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

Chapter 1

Introduction

Software and technologies are rapidly evolving with time and so programs are being more challenging day by day. More often the challenges are similar in nature. The swift progress of technology and frameworks are making people enthusiastic about software development. The rapid evolution of the software development industry necessitates the formation of a community to share knowledge and expertise with each other. A cooperation can save lots of effort and time. Q&A Platforms are the outcome of such necessity, which eventually results in the formation of a community for sharing knowledge. In the community, users share approaches, skills and techniques among themselves to solve various challenges. Among all other software development question-answer platforms of present decade, Stack Overflow is the largest and most renowned one [1]. It is a flagship site of Stack Exchange Network, created by Jeff Atwood and Joel Spolsky in 2008. With an aim to help one another with technical knowledge and wisdom Stack Overflow stated its journey and today it has created the largest knowledge-base in programming by accumulating all the challenging tasks and resolution of that. Since its inception, a total of 16.5 million people have registered on the site, with an average of 3,370 new members enrolling every day and making around 11,203 posts on a daily basis[1] (based on a query run in August 2021). The massive repository of 21 million questions and 31 million answers in Stack Overflow [2] is the result of the contribution of 16.5 million registered users. The daily inclusion of massive number of user and contribution of such big amounts of posts in the platform is the clear indicator of its popularity.

Stack Overflow has made this extensive dataset public [3][2] in the spirit of sharing and helping each other. The accessibility of this massive dataset has brought about a number of researches on this platform for more potential improvement and finding out the anomalies of the platform as a goal of being a better and more user friendly Q&A site.

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

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

Researches on various domains like community evolution [2,4], post analysis [5], code snippets [6], reputation [7], tags [8], badges [9], current trend [10] etc. has enriched the platform by suggesting steps to improvement. The study on community evolution of Stack Overflow depicts the evolution nature of Stack Overflow community, how the rapid expansion of this community collaborate with the goal and purpose of Stack Overflow, how user trends change over time, as well the effect of these changes on the community activities. The study on post analysis is evaluating users' posts to find patterns, user habits, community trends, etc. As posts are the most vital part of any networking website, analysis of posts gives us insights into the vision and purpose of the platform, how close are the users to the visions as well as the interaction between various users. A code snippet is a minor portion of source code, machine code, or text that can be reused and modified independently. Code snippets are used for further clarity in a question asked by any user. The study on code snippets analysis helps in categorizing questions topic-wise, as well as finding technology landscapes in Stack Overflow. Reputation is a rough estimate of how much the community trusts a user. It is acquired by persuading peers with programming knowledge. Reputation analysis helps in finding prominent users, studies privileges that come along with higher reputation, and evaluates the gravity of the privileges. Tags in Stack Overflow are predicted based on the questions asked by users. These tags help in categorizing posts based on the framework, programming language etc. Tags are essential in finding technology landscapes and as well interrelation between languages and frameworks. Badges are rewarded based on different milestones set by SO. These badges are offered for a diverse set of activities, so researchers often analyze them to find out user behaviour patterns (i.e. lurkers, active posters etc). The study on current trend finds and predicts which languages, frameworks, tech-stack etc. are popular currently. This field also focuses on user hospitality and interaction. However, not many studies are found on the hospitable environment of the community for its new users.

With the exponential growth of the community, the atmosphere and environment of the forum becomes the prime concern because any hostile nature of the community may turn off the zeal of participation. This hinders the lively ambience of the platform. However, the related studies intimate the existence of unwelcoming environments especially to the new users. László Tóth et. al. [11] showed in their study that frustrations creates among users with less experience due to the obscurity in closing questions. It eventually causes the community to become hostile and unsupportive, especially to new users. Antragama Ewa Abbas in his research [12] presented unanswered questions, negative feedback and deleted questions as the root of a massive discouraging impact that questions the healthy environment of the community and shrunken the involvement of new users. According to Rogier Slag et. al. [13], a remarkable user of

47% post only once and never come back to the community. As probable reasons, new user's posts get removed more often in addition to not receiving responses to their questions at a higher rate. Many researches on user participation [14, 15] imply the feeling of Stack Overflow environment to be hostile and unsupportive to new users. The gradual posts of renowned blog sites, official surveys, posts of meta stack exchange and official blogs of stack overflow community itself vocalizes the continual nature of the 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"*

A deep investigation can illustrate the reasons and help taking decisions to resolve the burning issue.

1.1 Research Questions

We address this issue in our study, by validating whether the unwelcoming nature of Stack Overflow is a reality and if so, investigate the probable reasons for new users facing such a hostile environment. To reach our goal, we first distinguish a distinct group of new users from the total users and classify them as *"neophytes"*. For validating our conjectures and fully comprehending the situations of neophytes in Stack Overflow, this study addresses two 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.

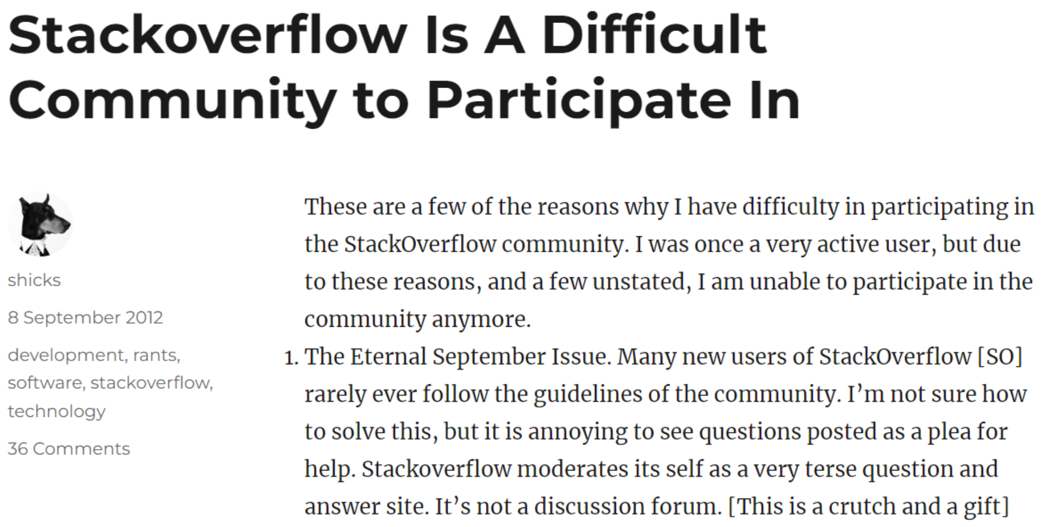
2. **RQ-2: What are the potential reasons for neophytes facing hurdles while collaborating in Stack Overflow?**
   There could be several potential reasons for which neophytes are facing hurdles. Identifying those reasons will help to understand the unwelcoming nature of the platform and provide insight towards solving the problem.

By answering RQ-1, we validate the problem of the unwelcoming environment of SO, specially to the neophytes. We find a number of potential reasons, including posts being deleted, closed, posting duplicated questions or answers, community rules violation etc. for facing a hostile environment by answering RQ-2

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

1.2 Motivation

Since the beginning, the unwelcoming community has been a buzzing issue. The negative feedbacks [16], offensive language [17], ambiguous closure of posts [11] make the platform hostile, specially affecting the new users. The issue's ongoing nature is also reflected in the gradual posts of well-known blog sites, official surveys, Meta Stack Exchange posts, and official blogs of the Stack Overflow community. A blog post of *the exception catcher* (Fig:1.1)[4] claims Stack Overflow as a difficult community for participation by observing the frequent downvoting tendency to a post.



Stackoverflow Is A Difficult Community to Participate In

shicks

8 September 2012

development, rants, software, stackoverflow, technology

36 Comments

These are a few of the reasons why I have difficulty in participating in the StackOverflow community. I was once a very active user, but due to these reasons, and a few unstated, I am unable to participate in the community anymore.

1. The Eternal September Issue. Many new users of StackOverflow [SO] rarely ever follow the guidelines of the community. I'm not sure how to solve this, but it is annoying to see questions posted as a plea for help. Stackoverflow moderates its self as a very terse question and answer site. It's not a discussion forum. [This is a crutch and a gift]

Figure 1.1: Blog by ExceptionCatcher

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[5] of Meta Stack Exchange on *"new-user"* tag urges the community to be supportive to them. It has been viewed 45 thousand times and received 1828 upvotes (according to August 2021). This clearly depicts that the community is not welcoming enough for new users (Fig:1.2).

Furthermore, a qualitative accumulation of evidence by Slegers[6] provides a verdict on the hatred nature of Stack Overflow whose major point is titled as *"Stack Overflow hates new users"*. The evidence is given in Fig:1.3.

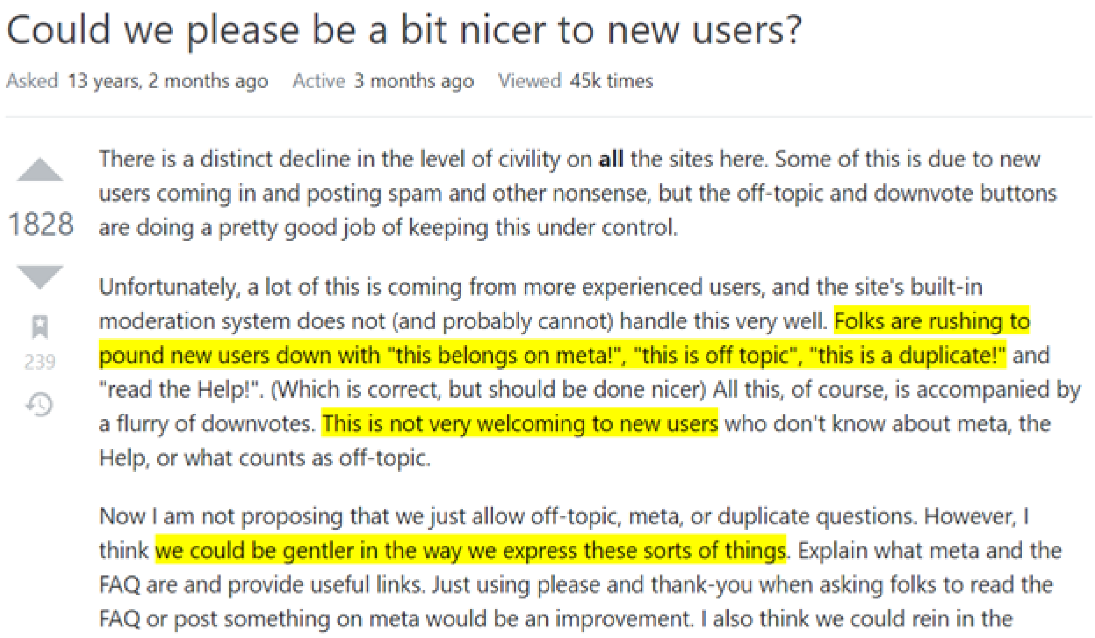
The former Executive Vice-President (EVP) of culture and experience of Stack Overflow is also vocal about the issue and asks for the prompt change of the situation (Fig:1.4)[7]. The following quotes from his discussion depicts the issue properly-

[4] https://theexceptioncatcher.com/blog/2012/09/

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

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

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



Could we please be a bit nicer to new users?

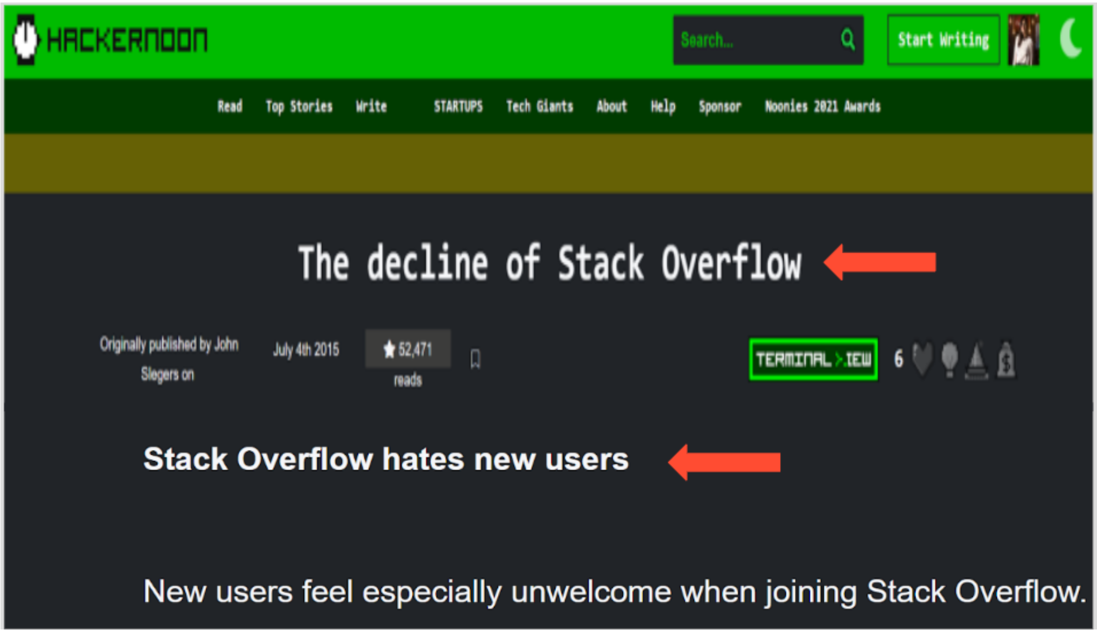
Asked 13 years, 2 months ago Active 3 months ago Viewed 45k times

There is a distinct decline in the level of civility on **all** the sites here. Some of this is due to new users coming in and posting spam and other nonsense, but the off-topic and downvote buttons are doing a pretty good job of keeping this under control.

Unfortunately, a lot of this is coming from more experienced users, and the site's built-in moderation system does not (and probably cannot) handle this very well. Folks are rushing to pound new users down with "this belongs on meta!", "this is off topic", "this is a duplicate!" and "read the Help!". (Which is correct, but should be done nicer) All this, of course, is accompanied by a flurry of downvotes. This is not very welcoming to new users who don't know about meta, the Help, or what counts as off-topic.

Now I am not proposing that we just allow off-topic, meta, or duplicate questions. However, I think we could be gentler in the way we express these sorts of things. Explain what meta and the FAQ are and provide useful links. Just using please and thank-you when asking folks to read the FAQ or post something on meta would be an improvement. I also think we could rein in the

Figure 1.2: Most Upvoted post of Meta Stack Exchange on "new-user" tag



The decline of Stack Overflow

Originally published by John Slegers on July 4th 2015 52,471 reads

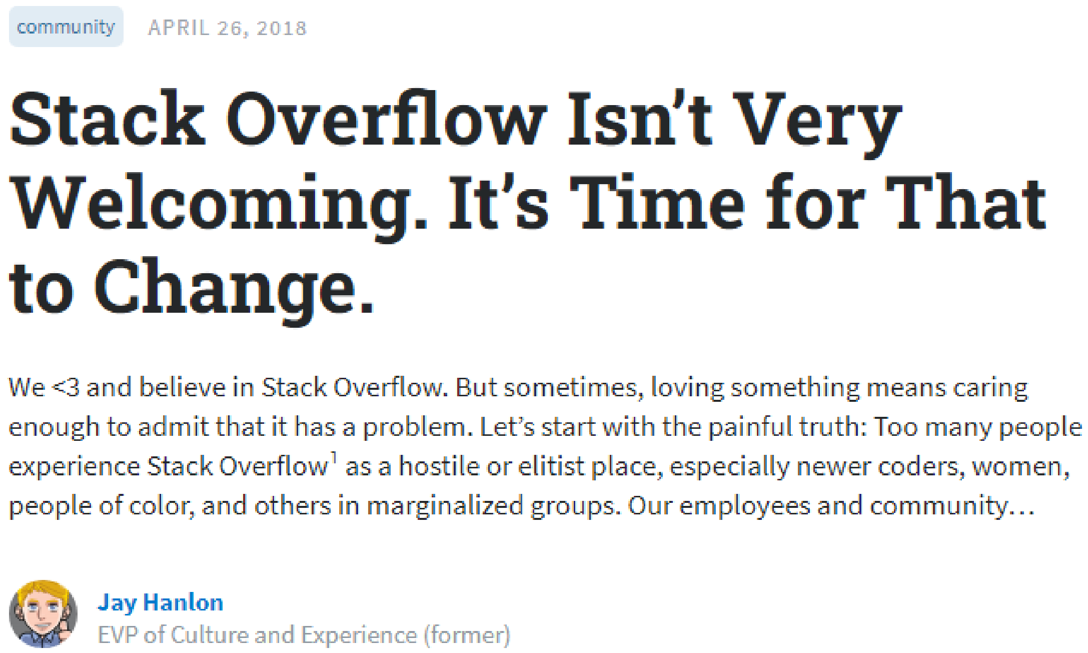
Stack Overflow hates new users

New users feel especially unwelcome when joining Stack Overflow.

Figure 1.3: Qualitative accumulation of evidence by Slegers

> *"Too many people experience Stack Overflow as a hostile or elitist place, especially newer coders".*

The hostile nature of Stack Overflow is not a problem of recent days, rather the situation has been prevailing since long ago and no improvement is reflected according to the site satisfaction survey of the community of 2019. The site satisfaction survey[3] is conducted by Stack Overflow community to obtain the insights about user pain points when using the site, is depicted in Fig:1.5 .



community APRIL 26, 2018

Stack Overflow Isn't Very Welcoming. It's Time for That to Change.

We <3 and believe in Stack Overflow. But sometimes, loving something means caring enough to admit that it has a problem. Let's start with the painful truth: Too many people experience Stack Overflow[1] as a hostile or elitist place, especially newer coders, women, people of color, and others in marginalized groups. Our employees and community…

Jay Hanlon
EVP of Culture and Experience (former)

Figure 1.4: Blog of Jay Hanlon

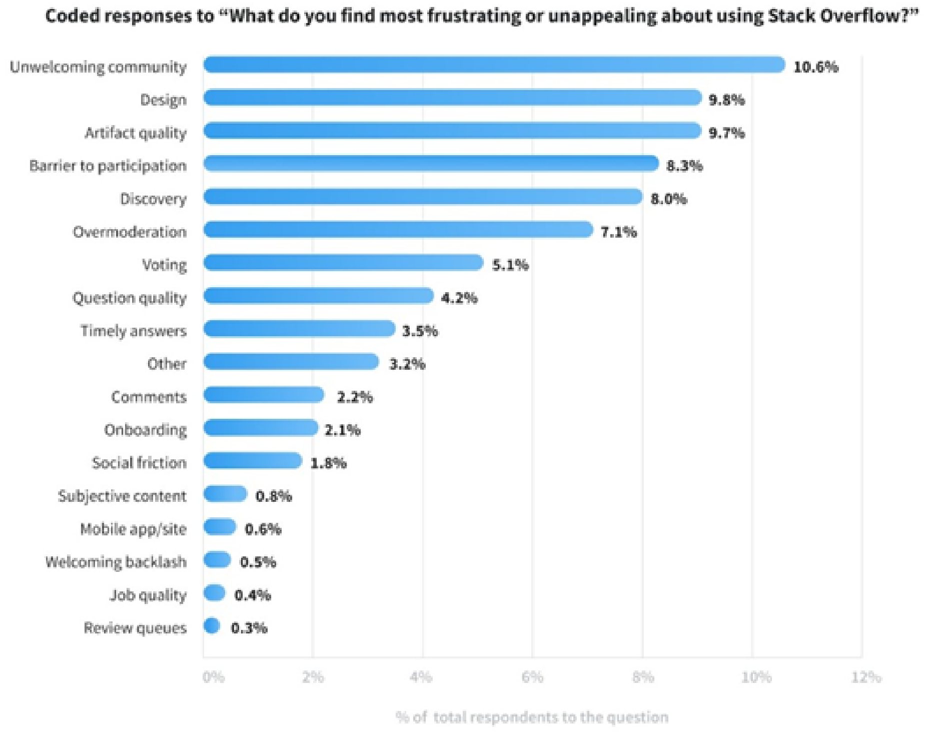


Figure 1.5: Site Satisfaction Survey 2019

In this survey, the Stack Overflow team asked users to answer the following question:

> *"What do you find most frustrating or unappealing about using Stack Overflow?"*

2,942 users gave response on that question. A perception of an unwelcoming community was the top thing that people found most frustrating or unappealing about Stack Overflow. About 10.6% of responses found **unwelcoming community** is the most frustrating or unappealing about using Stack Overflow.

According to developer survey of 2019[8] and 2020[9] we can clearly see the same scenario. The survey of 2019, which is depicted in (Fig:1.6) 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 (Fig:1.7) shows 70.6% vote.

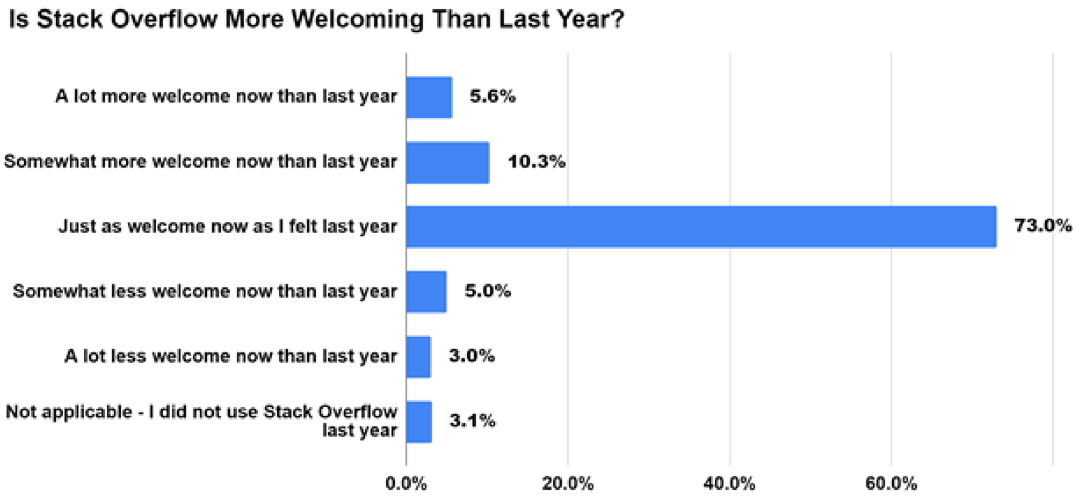


Figure 1.6: Developer Survey 2019

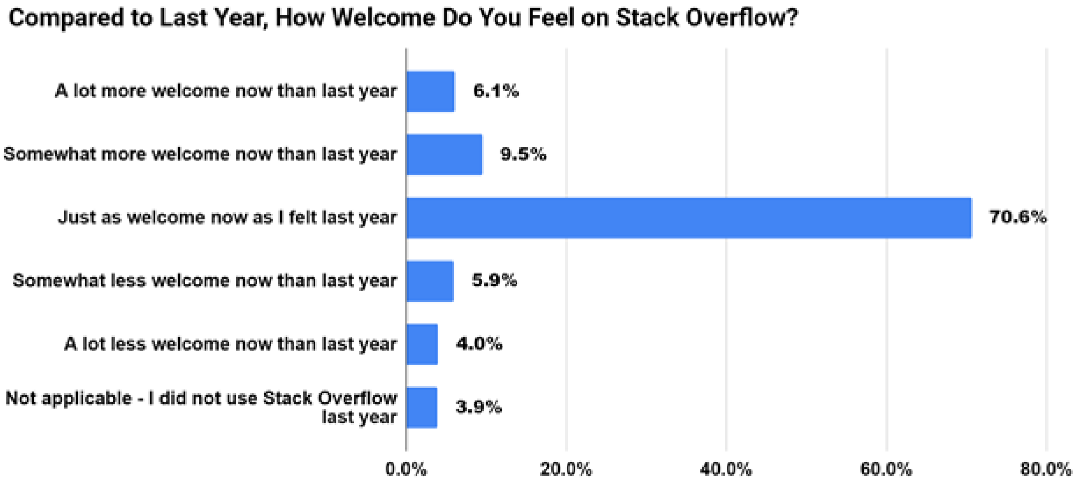


Figure 1.7: Developer Survey 2020

[8] https://insights.stackoverflow.com/survey/2019#community
[9] https://insights.stackoverflow.com/survey/2020

The user-base claimed to be undergoing through the issue is huge. 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.

Chapter 2

Related Work

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

2.0.1 Analysis of Stack Overflow Posts

"Post" analysis is one of the richest domains with research studies from the very beginning of Stack Overflow. Haifa Alharthi *et al.* [18] presented an approach to estimate question scores based on several factors. Duplicate questions are one of such factors. Duplicate questions are mostly posted by users with limited experience, of which duplicate answers contain distinctive information to help. The length of the question's code, approved response score, number of tags, and count of views, comments, and answers are all statistically significant factors in question scores, according to their study. Their findings help community-based Q&A sites improve the content of their collective knowledge. Duplicate questions are mostly posted by users with limited experience, of which duplicate answers contain distinctive information to help askers [19]. There are suggestions provided by studies to amend policies regarding handling duplicate posts in Stack Overflow to ensure better benefits for the overall community. According to Ripon K. Saha *et al.* [20] the indifference of the community to specific tags keeps the questions unanswered. The primary findings from their study indicate that there is a significant portions of the unanswered questions remain that way because they fail to catch the interest of the larger community.With the advancement of technology, the number of obsolete answers are increasing. Specific tags are also vulnerable to get obsolete answers and they are already obsolete at the time of posting [21]. A study by Zhang *et al.* [21] stressed on developing techniques to involve the community to be more mindful in terms of both asking and handling answers to reduce obsolete answers. Sarah Nadi *et al.* [22] crafted 4 techniques in their study for finding essential sentences to navigate through answers in Stack Overflow. By comparing the tech-

niques they proved that not any particular approach is always successful. To identify if a user can answer a new question, Morakot Choetkiertikul *et al.* [23] developed a predictive model in their study that takes the question topic and user reputation into account. The low reputed users are mostly new users and are accused to rely on intrinsic factors (answerer's reputation, representation of answer etc.) only to identify answer quality. However, the notion is proved wrong [24]. Tóth *et al.* [11] recently discussed the rivalry between the amount and quality of queries, as well as difficulties with the site's professionalism. They outlined the reasons for closing a post into 5 categories along with reporting their concern on ambiguous closing of posts. The ambiguous nature frustrates and hurts the users specially the new users and make the environment feel unwelcoming 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.

2.0.2 Analysis of Comments

Analysis of *"Comment"* is another important realm of research in Stack Overflow. Studies like categorizing the comments indicates how the comments help in learning and increasing skills [25]. One of the recent studies on SO investigated how the platform manages comments and claim that 97.3% answers are within the hidden comments section [26]. Analyzing these comments can provide with insights on gender hospitality in Stack Overflow [27]. A study on norm violations in SO shows that its comments are offensive and unwelcoming by presenting a taxonomy of norms that are violated [17]. Abhishek Soni *et al.* [28] built a system in their study to update the obsolete answers by scrutinizing the comments. It is an amazing initiative to resolve obsolete answer problems. A general study of comments by Wenhan Zhu *et al.* [29] found an inverse correlation between comments in a question and time required to answer it. Comment is a vital aspect to understand the environment and culture of a platform. Unfortunately, the domain still requires research in these, specially addressing the situation of new users.

2.0.3 Dynamics of User behaviour

Various studies on user badge, reputation, participation have been making the Stack Overflow *"User"* domain enriched since the very dawn of its establishment. A study by Stav Yanovsky *et al.* [30] discussed the association of user contribution and behavior with achievement of badges. Furthermore, the authors falsified the claim of increasing user participation along with the achievement of badges. Andrew Marder *et al.* [31] aligned with the previous study and suggested an alternative approach towards users.

A much needed contribution for the new users is the research of Amiangshu Bosu *et al.* [7], where they provided guidance to new users on enhancing their reputation swiftly. Laura MacLeod *et al.* [32] inspected the correlation between users reputation and heterogeneity of tags of their contribution. Additionally, they depicted a poor community structure.

2.0.4 Correlation Between User Behaviour and Reputation

There is a high association of involvement habits of individuals with high and low reputation. It is a fact that extremely high-reputation users are the dominant source of replies, particularly high-quality responses [15]. On the contrary, low-reputation users ask a bulk of questions on the site. Users with complete profiles have relatively higher reputation and also they post higher quality contents [14]. Furtado *et al.* [33] evaluated all contributors' activity from a huge site called *"Super User"*, and discovered nine behavioral profiles that categorize people based on the quality and quantity of their contributions. The profiles discovered improve general understanding of how Q&A sites operate, and knowing these traits can help site management.

2.0.5 Concerns regarding Stack Overflow Environment

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 [34]. 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 [13]. 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 [12] 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 [35]. 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 [16].

While all the studies mentioned above contribute to significant aspects of user reputation, badge, participation and community environment in Stack Overflow, there is no 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.

Chapter 3

Methodology For Empirical Study

3.1 Methodology Overview

At the very beginning of our work, the Stack Overflow online dataset of 2020 has been selected for the research. The study needed to specify a group of users for consistent analysis and such users are termed as neophytes. The neophytes are then defined based on their newness and contribution to the platform. From the dataset, we extract neophytes' data. Followed by the neophytes' data extraction a qualitative as well as quantitative analyses are performed to answer the RQ-1 and RQ-2. The qualitative analysis is performed considering posts, blogs, surveys and meta discussions 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 sections and subsections describe the methodology with further details along with depicting an overview of it in Figure 3.1.

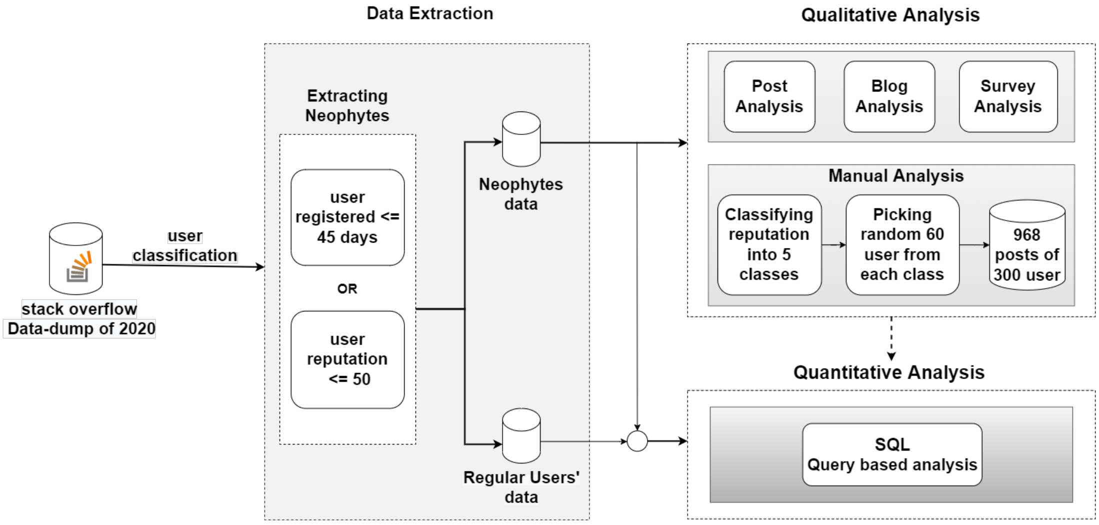


Figure 3.1: An overview of methodology.

3.2 Dataset Extraction

The first task to address the issue is to select a specific dataset on which further investigation can be done to answer our research questions. A specific dataset helps keeping consistency in the analysis.

3.2.1 Selecting Offline Database from Stack Exchange Archive

Stack Exchange archive[10] provides the official Stack Overflow data. It stores all the previous versions of databases as well, starting from the very beginning of Stack Overflow's journey i.e. 2008. A new version of database is linked to the site every year. Moreover, the data of a month are added to the yearly database on a monthly basis.

Each of the categories of data (post, user, comment, flag etc.) are classified as a dataset. All the datasets are zipped as 7zip. Unzipping a dataset provides the data as xml format. To use the data one has to convert the xml data into mdf format which is the acceptable database file format in SQL Server.

Because the offline databases are much more stable and convenient to work with, we chose to work with the offline database of Stack Overflow. However, the main challenge to work with offline database is to convert this huge file into mdf format. Often a file ranges from 10GB to 80GB. So, it requires a high-power server computer to convert and store such huge data files. Furthermore, the offline database lacks some important datasets like *CloseAsOffTopicReasonTypes*, *CloseReasonTypes*, *FlagTypes*, *PendingFlags*, *SuggestedEdit*, *PostFeedback* etc. The datasets are vital for our analysis as these are the concrete indicators of hurdles of SO users.

3.2.2 Shifting to Online Database (Stack Exchange Data Explorer)

Because of the missing of some vital datasets in offline database we shifted our entire work to the online version of the Stack Overflow database i.e. *Stack Exchange Data Explorer*[11]. The Stack Exchange Data Explorer is a tool for performing arbitrary SQL queries against data from the various question and answer sites in the Stack Exchange Network. It provides easy web-based access to the latest monthly Stack Exchange data dumps. The site also provide a compilation environment to query on the available data of Stack Exchange Network.

The advantage of online database over the offline one is that it provides all data except only from deleted data. Stack Overflow never discloses any data on deleted posts. The

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

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

deleted posts are removed from the site by the post owner. This is why the data are considered as private to the owners and Stack Overflow keeps them confidential. Other than the data of deleted posts, all other datasets are available.

The only drawback of the online database is that it is constantly changing on the monthly basis. A user cannot fix up a database to work on. The dynamic nature of the database causes the query results alter on a monthly basis. However, we discovered a way out of this problem by fixing up the data range within a time frame of 2020. With every query we considered only those data that has been posted within the range of 1st January 2020 to 31st December 2020. This causes the data to behave as static offline database.

3.2.3 Selecting Timeframe of 2020

For consistent analysis on the data, we had to consider a dataset of a specific timeframe. The latest database available during the search work was of 2021. However, the database of 2021 was not completed by that time.

Working with an incomplete database has a high chance of creating inconsistency in the result. Moreover, we aimed at working with the data of an entire year. This was not possible with the 2021 database, as full database of 2021 would not be available until January of 2022.

The one-year timeframe seems to be quite enough as most other works in this domain used data of timeframe ranging from six months to a year. So, we decided to work with the database of previous year (2020). The database of 2020 was complete and much more stable that that of 2021 by that time.

3.3 Defining Neophytes

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 [1]. 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”. There are two constraints – one of being new to the community and another of having less contribution to the platform. For being a neophyte, a user has to satisfy one of the two constraints.

3.3.1 Defining Newness to the Community

For defining new members to the community, we can look at the definition of "new user" according to Stack Overflow. Stack Overflow terms a user registered to their site not more than 45 days as a "new user".

Although it is an acceptable indication of newly joined users to the platform, it does not specify anything on their contribution to the platform. The contribution of a user on the platform is well understood by their reputation. According to Stack Overflow –

> *"Reputation is a rough measurement of how much the community trusts you; it is earned by convincing your peers that you know what you're talking about."*

So, we add a reputation boundary to understand the contribution level of a user.

3.3.2 Adding the Reputation Boundary

As we cannot ensure the level of contribution only by restricting the definition of neophyte with duration of registration, 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. in their research *"One-day flies on stackoverflow - why the vast majority of stackoverflow users only posts once"* [13] worked with 38 reputations 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. 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 reputations. So, to ensure an impactful presence of users in SO, 50 reputation is chosen over 38.

Therefore, the final definition of neophyte is –

> *"Neophytes are those groups of users who are either registered in Stack Overflow within the last 45 days or have a reputation of less than or equal to 50."*

If one of the conditions gets satisfied for a user, that user will be considered as a neophyte. Everyone other than neophytes in Stack Overflow is specified as "regular users" throughout the work. The algorithm to separate new users from regular users is presented in Algorithm 1 -

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 $(user.reputation \leq 50)$ $or$ $(user.registration\_day \leq 45)$ then
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 reputation boundary or the registration date within last 45 days is checked in line-4. One fulfilling any of the constraints is added to the neophytes list. According to our definition, 89.9% (14,897,718) of the total users of Stack Overflow data dump 2020 are neophytes.

3.4 Dataset Description

For our study, we focus on a specific group of users to investigate the attitude of SO community towards them. According to our definition, 89.9% (14,897,718) of total users of Stack Overflow data dump 2020 are neophytes. The detailed dataset description is depicted in the table below –

Total	Total Users	Total Posts
	14,897,718 (14.9 Million)	4,456,062 (4.5 Million)
Neophyte	Total Neophyte	Neophytes' Total Post
	12,577,534 (12.6 Million)	1,161,701 (1.2 Million)
Regular User	Total Regular User	Regular Users' Total Post
	2,320,184 (2.3 Million)	3,294,361 (3.3 Million)

Table 3.1: Dataset Description.

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. So, a regular user on an average post more than 7 posts whereas the average post count of a neophyte is below 2.

3.5 Methodology of RQ-1

RQ-1 validates whether neophytes actually face hurdles while collaborating in Stack Overflow. To obtain the research questions outcome we perform both qualitative and quantitative analysis. Further details of the analyses are described below-

3.5.1 Qualitative Analysis

The qualitative analysis comprises of post analysis, blog analysis[7], survey analysis[3], meta discussion analysis[5] and manual analysis.

Post Analysis

Neophytes' posts are taken into consideration for the analysis and to understand their situation in the platform. Posts are manually analyzed to inspect any issue. Post's quality, evaluation, comment, up and downvote are analyzed thoroughly. A number of posts are found to create quarrelsome situation where neophytes and regular users blame each other. Some of them contain rude comments as well.

Blog Analysis

The blog analysis includes both the official and unofficial blogs. The official blog of *"Jay Hanlon"*[7], former Executive Vice President (EVP) of culture and experience at Stack Overflow urges the need of changing Stack Overflow environment. Other than this *"Born Geek"*[12], *"The Exception Catcher"*[4], *"Hackeroon"*[6] posted their blog on declining friendly environment in Stack Overflow.

Survey Analysis

Every year Stack Overflow publishes their site related surveys. The site satisfaction survey[3] and the developer survey[9] are the prominent surveys. Both the surveys of various years have been analyzed to understand how the users feel regarding the platform's environment. The outcome shows the existence of a hostile environment for years.

[12] https://borngeek.com/2012/01/04/stack-overflow-hates-new-users/

Meta Discussion Analysis

Meta Stack Exchange is intended for bugs, features, and discussions that affect the whole Stack Exchange family of Q&A sites. Various meta posts question the environment of SO. Analysis on those meta posts has been conducted. The analysis of posts clearly depicts the division of two groups one blaming the other.

Manual Analysis (Post-based)

300 neophytes who 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.

An overview of the manual analysis process has been depicted in Figure 3.2. 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 randomly accumulated 300 neophytes posted a total of 968 posts. These 968 posts are analyzed to answer RQ-1. The outcome of manual analysis provides a convincing stat on neophytes facing hurdles while participating in SO. The result of manual analysis is presented in the result section of RQ-1.

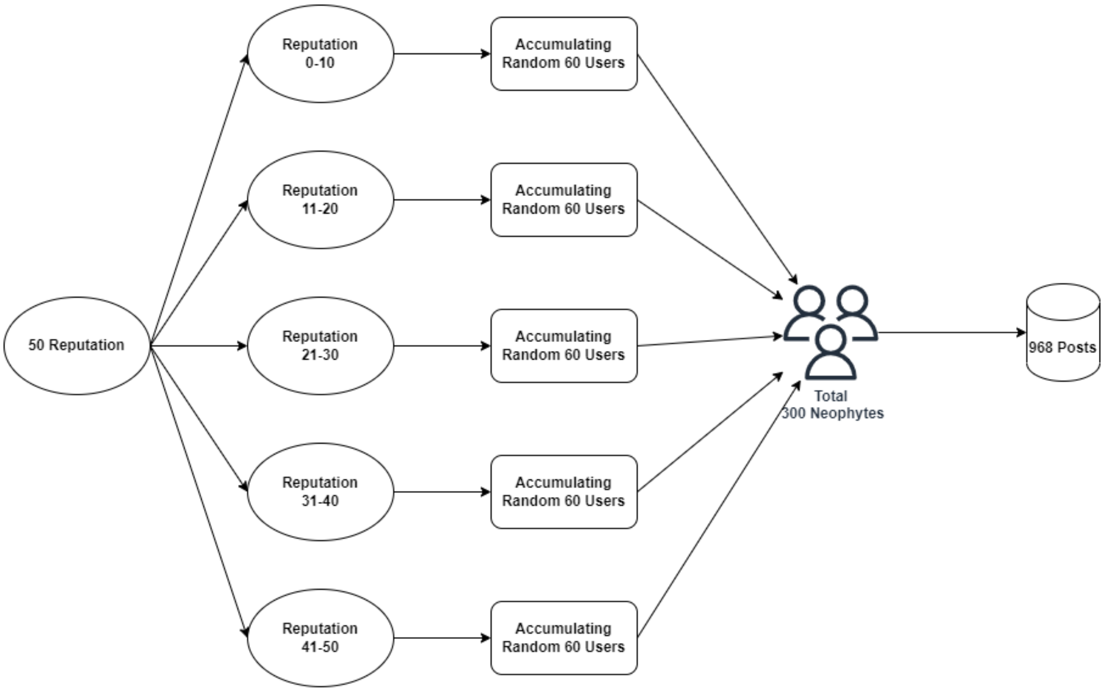

50 Reputation
Reputation 0-10
Reputation 11-20
Reputation 21-30
Reputation 31-40
Reputation 41-50
Accumulating Random 60 Users
Accumulating Random 60 Users
Accumulating Random 60 Users
Accumulating Random 60 Users
Accumulating Random 60 Users
Total 300 Neophytes
968 Posts

Figure 3.2: Manual Analysis Process.

The following factors have been considered while manually analyzing the posts -

- Whether post is downvoted or not.
- Whether post is closed or not.
- Whether posts received any response.
- Whether post is duplicate or not.
- If a downvoted post received any reason mentioned.
- If there is any rude comment mentioned in the post.

3.5.2 Quantitative Analysis

The quantitative analysis has been performed on the official online query site of Stack Exchange Network, **"Stack Exchange Data Explorer"**. For finding result of RQ-1, a number of queries are formulated. Some of the queries include -

- Comparing total posts with the number of posts of neophytes.
- Number of downvoted posts of neophytes.
- Whether first post of neophytes are downvoted.
- Whether neophytes continued participating on the platform after getting downvoted.
- Comparing average view count of neophytes' posts with respect to total posts.

The queries are formulated in a way to obtain the result of RQ-1 only.

3.6 Methodology of RQ-2

To answer second 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.

3.6.1 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 968 posts, the 300 neophytes' profiles have been considered as well to answer RQ-2.

Manual Analysis (Post-based)

The post-based manual analysis is performed with more detailed investigation. After analysis of each post an observation is put against that post. The observation depicts the reason of the situations the posts faced. It includes both welcoming and unwelcoming situations.

The point of view of this analysis has changed from whether posts are evidence of hostile environment to accumulating probable reasons for such unwelcoming environment. A list of reasons is accumulated which are commonly found among the posts.

Manual Analysis (Profile-based)

Profiles of each of the 300 neophytes are taken into consideration in this phase. 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.

3.6.2 Quantitative Analysis

To delicately investigate and verify the obtained list of reasons from the qualitative analysis, a query-based quantitative analysis is conducted. We have formulated numerous query-based questions and executed the queries in the online query site of Stack Overflow, the “Stack Exchange Data Explorer”. Some of the noteworthy ones include –

- Number of posts of neophytes that has got no response at all.
- Number of negatively scored posts where reason of down-vote is missing.
- Comparing number of duplicate posts of neophytes with respect to regular users.
- Comparing number of posts of neophytes those are closed with respect to the regular users.
- Number of neophytes stopped posting after their posts being responseless.
- Number of neophytes getting “informed” badges, “peer-pressure” badges. The "informed" badge represents that a user has gone through the entire tour page and knows the basics of SO rules and regulations. Whereas, the "peer-pressure" badge is given when a user delete his/her post after facing negative score.

Chapter 4

Result Analysis

The research questions along with the methodology described in the previous chapters give us a clear pictorial view of the overall issue, leading to the community thinking about taking effective and long term steps towards resolution.

Keeping consistency with the Methodology, the results are also represented based on the relevant research question

4.1 Result of RQ-1

A total of 968 posts were analyzed for this research question. Among them, 254 posts are given a score less than zero 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 4.1.

Among all the posts analyzed, almost 49% of the posts indicate neophytes facing some sort of difficulties. The difficulties include posts being closed, posts marked as duplicate, no response to posts, negative scored posts and so on. In our analysis. 254 posts are under the reason label "negative scored posts". 123 posts are categorized under "posts with no explanation or proper cause of getting the negative score". The category with 123 posts was not considered separately as it is a subset of the category with 254 posts. 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, individuals of all reputation levels made a total of 4,456,062 posts between January 2020 and December 2020. Users classified as neophytes made 1,161,701 posts among them. This means that neophytes accounted

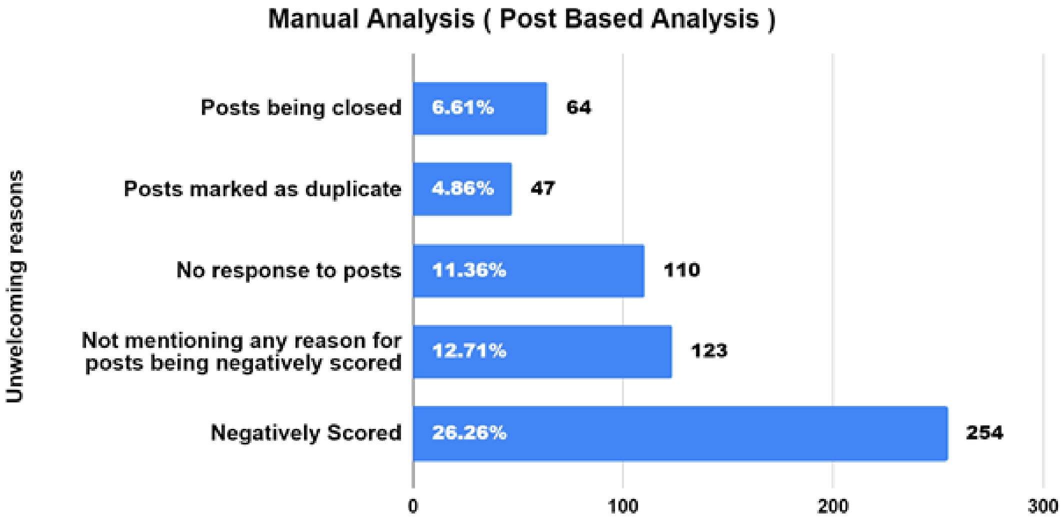


Figure 4.1: Post Based Manual Analysis.

for more than a quarter (26.07 percent) of the platform's total posts. This further emphasizes the fact that neophytes make up a sizable section of the community in terms of post contribution. A total of 108,568 posts, or 9.35 percent of all novice posts, receive a bad score. The percentage might be misleading as it is small and seems very normal. But the same number for the regular users comes down to 98,830 which is 3% of total regular users. 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.

4.2 Result of RQ-2

Among manual analysis of 300 neophytes' profiles depicted in Figure 4.2, total 77 neophytes obtained the *"Informed"* badge, a badge that is awarded to users who have visited the FAQ page (now known as "tour page") containing basic information about SO. This indicates that only 25.67% of neophytes undergo the entire tour page to gain knowledge on how Stack Overflow works. A total of 41 neophytes did not post further after their posts were negatively scored. these users cover 13.67% of our accumulated data-set. Moreover, 62 neophytes posted only once, defined as *"one-day-flies"* [13], 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.

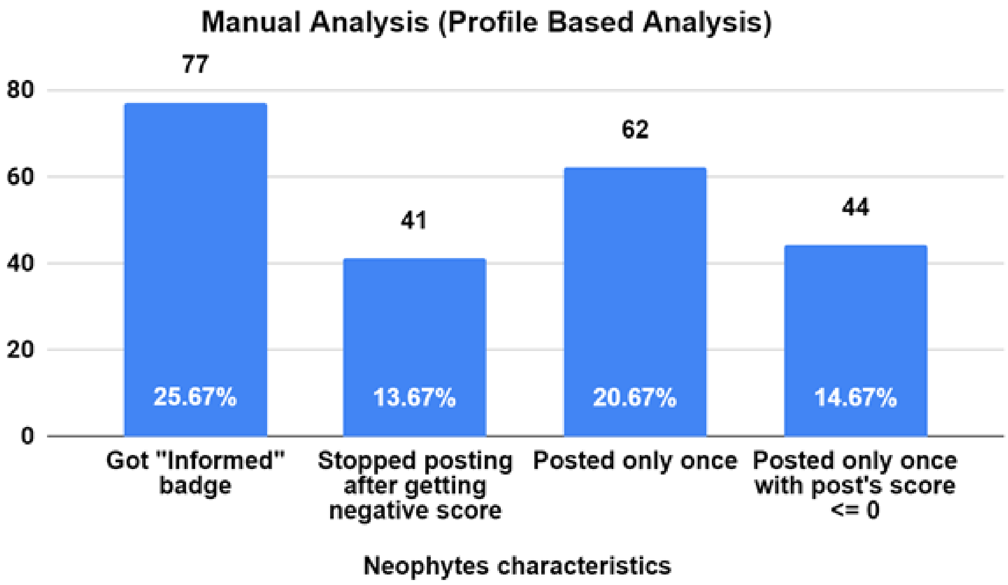


Figure 4.2: Profile Based Manual Analysis.

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 4.1 and Listing 4.2 are two of those queries.

```sql
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 4.1: Query to find posts marked as duplicate.

```sql
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 4.2: Query to find posts with no reason for being negatively scored.

The SQL in Listing 4.1 provides a query to determine the number of neophyte duplicate posts. If a post has the property *linkTypeId* equal to 3, it is considered a duplicate of a previous post that is substantially similar to it. The SQL in Listing 4.2 denotes a query to determine the number of posts with no explanation or legitimate cause for receiving a negative score. A post with a negative score is deemed to be adversely scored without cause if it has no comments, pending flags, recommended edits, or is being closed. Because each closed post has a reason, the closed post constraint is also incorporated 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

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%)

Table 4.1: Comparison of total posts and neophytes posts.

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 4.1 shows a comparison between total posts vs neophytes' posts. Whereas, Figure 4.3 depicts the stats of several reasons from the quantitative analysis.

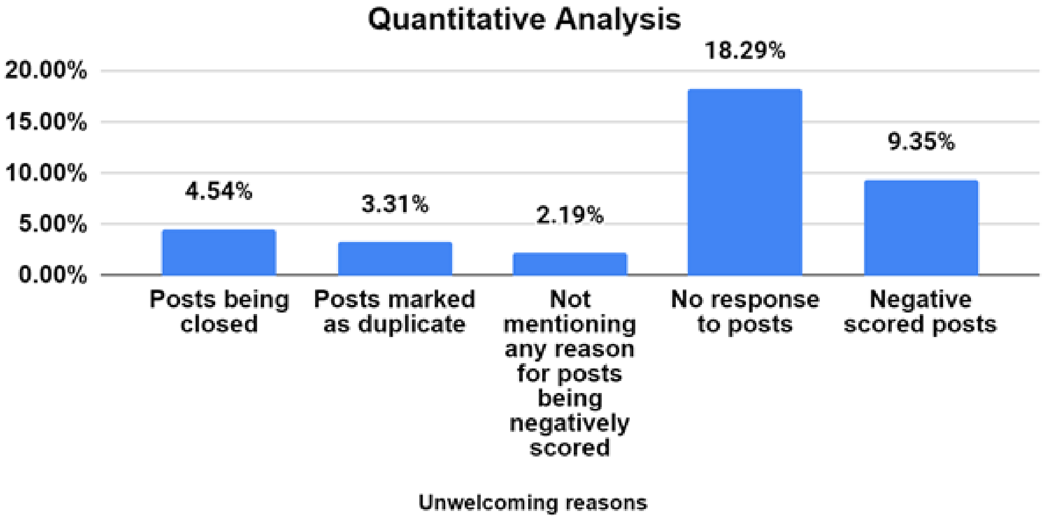


Figure 4.3: Quantitative Analysis

4.2.1 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 4.1. Figure 4.3 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.

4.2.2 Posts Marked as Duplicate

Our analysis in Table 4.1 states that, in 2020 total 78,652 posts were marked as duplicate where 38,508 (48.96%) posts belong to neophytes. According to Figure 4.3, 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, Durham Abric *et. al* [19], 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.

4.2.3 Not Mentioning Any Reason for Posts being Negatively Scored

From Table 4.1, 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 4.3. 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.

4.2.4 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 response-less which is presented in Figure 4.3 and Table 4.1. 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.

4.2.5 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.

4.2.6 Deletion of Posts

According to Rogier Slag *et. al* [13], 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. Antragama Ewa Abbas discussed *"Deleted Questions"* as one of the significant factors for people not participating in SO [12].

As Stack Overflow keeps all the information related to deletion of posts private[13], 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.

4.2.7 Moderation Without Proper Reasoning

In SO, users get responses within a very short period of time, typically within 21 minutes[14]. 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.

4.2.8 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

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

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

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. This ultimately leads them to leave the community.

4.2.9 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.

Chapter 5

Recommendation and Conclusion

5.1 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.

5.1.1 Pre-post Prediction of Whether a Post will be Closed or Not

For closed posts, SO is recommended to use a pre-post automated prediction tool. The tool will predict whether a post will be closed or not before the post is published. This will lead users identify if their posts are going to be closed in future.

Only predicting the chance of closure of posts will not be sufficient. Because, this does not inform users the problem in the post that cause it to be closed. So the tool also need to 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.

5.1.2 Imposing Moderators to Mention Reason Behind Moderation

Moderation is important to ensure the quality of Stack Overflow platform. However, a moderation without providing reasons does not let users understand their fault. In fact, it leads them to frustration. One of such moderation is downvoting a post.

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.

5.1.3 Identifying Rude Comments Before They are Published

Rude language spreads negativity to the platform. SO has their own bot that detects rude comments. But it can only do so after the comment has been published. By the time, the bot can detect rude language and take actions, it comes to the notice of users and the damage is done.

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.

5.1.4 Identify the Responseless Posts

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.

However, finding answers to responseless posts might be hard. Rather, a pre-post prediction tool can be developed to predict the probable response time of a post. Posts with longer response time than a threshold will be certain to not get response.

5.1.5 Proper Assessment to Ensure Users are Well Acquainted to SO Rules

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.

Stack Overflow should impose an assessment that validates the acquaintance of neophytes to SO rules and norms before participating to the platform.

5.2 Threats to Validity

5.2.1 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.

5.2.2 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.

5.3 Future Work and 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 to build a welcoming environment by realizing current practices towards neophytes and creating awareness to 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

[1] A. May, J. Wachs, and A. Hannák, “Gender differences in participation and reward on stack overflow,” *Empirical Software Engineering*, vol. 24, no. 4, pp. 1997–2019, 2019.

[2] I. Moutidis and H. T. Williams, “Community evolution on stack overflow,” *Plos one*, vol. 16, no. 6, p. e0253010, 2021.

[3] B. Bazelli, A. Hindle, and E. Stroulia, “On the personality traits of stackoverflow users,” in *2013 IEEE International Conference on Software Maintenance*, 2013, pp. 460–463.

[4] G. Blanco, R. Pérez-López, F. Fdez-Riverola, and A. M. G. Lourenço, “Understanding the social evolution of the java community in stack overflow: A 10-year study of developer interactions,” *Future Generation Computer Systems*, vol. 105, pp. 446–454, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19311884

[5] S. Baltes, L. Dumani, C. Treude, and S. Diehl, “The evolution of stack overflow posts: Reconstruction and analysis,” 2018.

[6] S. Baltes, C. Treude, and S. Diehl, “Sotorrent: Studying the origin, evolution, and usage of stack overflow code snippets,” in *Proceedings of the 16th International Conference on Mining Software Repositories*, ser. MSR ’19. IEEE Press, 2019, p. 191–194. [Online]. Available: https://doi.org/10.1109/MSR.2019.00038

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

[8] A. K. Saha, R. K. Saha, and K. A. Schneider, “A discriminative model approach for suggesting tags automatically for stack overflow questions,” in *2013 10th Working Conference on Mining Software Repositories (MSR)*, 2013, pp. 73–76.

[9] A. Halavais, K. H. Kwon, S. Havener, and J. Striker, “Badges of friendship: Social influence and badge acquisition on stack overflow,” in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 1607–1615.

[10] I. K. Villanes, S. M. Ascate, J. Gomes, and A. C. Dias-Neto, “What are software engineers asking about android testing on stack overflow?” in *Proceedings of the 31st Brazilian Symposium on Software Engineering*, ser. SBES’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 104–113. [Online]. Available: https://doi.org/10.1145/3131151.3131157

[11] L. Tóth, B. Nagy, T. Gyimóthy, and L. Vidács, “Why will my question be closed? nlp-based pre-submission 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)*, 2020, pp. 45–48.

[12] A. E. Abbas, “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)*, vol. 7, 2019, pp. 1–5.

[13] 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*, 2015, pp. 458–461.

[14] I. Adaji and J. Vassileva, “Towards understanding user participation in stack overflow using profile data,” in *International Conference on Social Informatics*. Springer, 2016, pp. 3–13.

[15] 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)*, 2013, pp. 886–893.

[16] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “How community feedback shapes user behavior,” in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

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

[18] H. Alharthi, D. Outioua, and O. Baysal, “Predicting questions’ scores on stack overflow,” in *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering*, ser. CSI-SE ’16. New York, NY,

USA: Association for Computing Machinery, 2016, p. 1–7. [Online]. Available: https://doi.org/10.1145/2897659.2897661

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

[20] R. K. Saha, A. K. Saha, and D. E. Perry, "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*, ser. ESEC/FSE 2013. New York, NY, USA: Association for Computing Machinery, 2013, p. 663–666. [Online]. Available: https://doi.org/10.1145/2491411.2494585

[21] H. Zhang, S. Wang, T.-H. Chen, Y. Zou, and A. E. Hassan, "An empirical study of obsolete answers on stack overflow," *IEEE Transactions on Software Engineering*, vol. 47, no. 4, pp. 850–862, 2021.

[22] S. Nadi and C. Treude, "Essential sentences for navigating stack overflow answers," in *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2020, pp. 229–239.

[23] M. Choetkiertikul, D. Avery, H. K. Dam, T. Tran, and A. Ghose, "Who will answer my question on stack overflow?" in *2015 24th Australasian Software Engineering Conference*, 2015, pp. 155–164.

[24] K. Hart and A. Sarma, "Perceptions of answer quality in an online technical question and answer forum," in *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*, ser. CHASE 2014. New York, NY, USA: Association for Computing Machinery, 2014, p. 103–106. [Online]. Available: https://doi.org/10.1145/2593702.2593703

[25] S. Sengupta and C. Haythornthwaite, "Learning with comments: An analysis of comments and community on stack overflow," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[26] H. Zhang, S. Wang, T.-H. P. Chen, and A. E. Hassan, "Are comments on stack overflow well organized for easy retrieval by developers?" *ACM Trans. Softw. Eng. Methodol.*, vol. 30, no. 2, Feb. 2021. [Online]. Available: https://doi.org/10.1145/3434279

[27] S. Brooke, "“condescending, rude, assholes”: Framing gender and hostility on stack overflow," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 172–180.

[28] A. Soni and S. Nadi, "Analyzing comment-induced updates on stack overflow," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 2019, pp. 220–224.

[29] W. Zhu, H. Zhang, A. E. Hassan, and M. W. Godfrey, "An empirical study of question discussions on stack overflow," *arXiv preprint arXiv:2109.13172*, 2021.

[30] S. Yanovsky, N. Hoernle, O. Lev, and K. Gal, "One size does not fit all: A study of badge behavior in stack overflow," *Journal of the Association for Information Science and Technology*, vol. 72, no. 3, pp. 331–345, 2021.

[31] A. Marder, "Stack overflow badges and user behavior: An econometric approach," in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 2015, pp. 450–453.

[32] L. MacLeod, "Reputation on stack exchange: Tag, you're it!" in *2014 28th International Conference on Advanced Information Networking and Applications Workshops*, 2014, pp. 670–674.

[33] A. Furtado, N. Oliveira, and N. Andrade, "A case study of contributor behavior in q&a site and tags: the importance of prominent profiles in community productivity," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, pp. 1–16, 2014.

[34] J. Cheriyan, B. T. R. Savarimuthu, and S. Cranefield, "Towards offensive language detection and reduction in four software engineering communities," in *Evaluation and Assessment in Software Engineering*, 2021, pp. 254–259.

[35] C. Lampe and E. Johnston, "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*, ser. GROUP '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 11–20. [Online]. Available: https://doi.org/10.1145/1099203.1099206