Analysis of 'Closed Questions' on Stack Overflow*

Md Meher Hassan Chowdhury
Dept of Computer Science
Lakehead university
Thunder bay, Canada
mchowdh8@lakeheadu.ca

***Abstract*—Stack Overflow is a Website hosting questions and answers on a wide range of topics in computer programming. Amongst all programming questions there are a lot that has no links with the field of questionnaires. These questions are usually closed by the moderator or experienced users. There are ten reasons why a question is closed i.e 'Exact Duplicate', 'Duplicate', 'Off-topic', 'Unclear what you're asking', 'Too broad','Primarily opinion-based', 'Off-topic','Subjective and argumentative', 'Not a real question', 'Too localized'. In this paper, I worked on analyzing and visualizing of ten years available data that contains over million closed question's.**

***Index Terms*—closed question, analyzing, visualization**

I. Introduction

Programmers most of the times finds difficulties solving their projects. Many of the times they don't get the prior knowledge's from their team and relies on some community who are willing to answer about the queries someone have. There are many question answering sites yahoo for example where many professionals willingly communicates about the questions someone asks. Stack overflow is considered the most popular Question and Answer forum in recent days. It's also free to use. A question need to be tagged with the subject and a body that determines the problem. There is a chatting section where any user can provide solutions or related ideas according to the problem. There are also voting system based on the level of question asked and also for the contribution made bu the community people who tries to solve the problem. Questions on Stack Overflow which do not fall into one of the pre-defined set of guidelines are marked 'closed' via a community-based voting system.

II. Research Contributions

The study provides the following contribution:

- Characterization of closed questions
- Analyzing the question contents
- Observations

III. RELATED WORK

Denzil Correa, Ashish Sureka proposed Analysis and Prediction of 'Closed Questions' on Stack Overflow where they worked with 4 years of publicly available data which contains 3.4 Million questions.They analyzed and characterized the complete set of 0.1 Million 'closed' questions. then used a machine learning framework and build a predictive model to identify a 'closed' question at the time of question creation. The context of this paper is the same for the first part of the related work. But the data I worked on in enriched from 2014 to 2018.

IV. 'CLOSED' QUESTIONS ON STACK OVERFLOW

A question is closed if it receives five close votes. A moderator can close a question immediately. The reasons for the questions are closed are:

- 1 = Exact Duplicate
- 2 = Off-topic
- 3 = Subjective and argumentative
- 4 = Not a real question
- 7 = Too localized
- 10 = General reference
- 20 = Noise or pointless (Meta sites only) Current close reasons:
- 101 = Duplicate
- 102 = Off-topic
- 103 = Unclear what you're asking
- 104 = Too broad
- 105 = Primarily opinion-based

V. CHARACTERIZATION STUDY OF 'CLOSED' QUESTIONS

In this part a characterization study of 'closed' questions on Stack Overflow is performed.

A. Dataset Description

The data set contains user, post history and votes with over million of data from 2008 to 2018. All the questions are closed in the dataset. The dataset contains 732847 number of closed questions from where the ratio of closed question for each year is calculated. The following table shows the distribution of yearly percentage along with the ratio of closed question over all the asked question until 2018.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Closed Questions	0.024289	0.146415	0.765508	3.459794	9.220342	15.210542	16.404925	13.302777	14.719171	14.988667	11.75757	732847 (5.11%)

Fig. 1. Percentage of Yearly closed questions.

Despite the presence of vibrant community and structured guidelines, users do post questions which are unfit for the website.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Closed votes	0.501238	2.617574	3.570545	7.0297	10.717822	16.930693	16.373762	11.200495	11.435644	11.998762	7.623762	16160.0

Fig. 2. Percentage of Yearly closed votes.

The ten categories for which a question is closed is shown on the following pie chart. It shows that Duplicate, off-

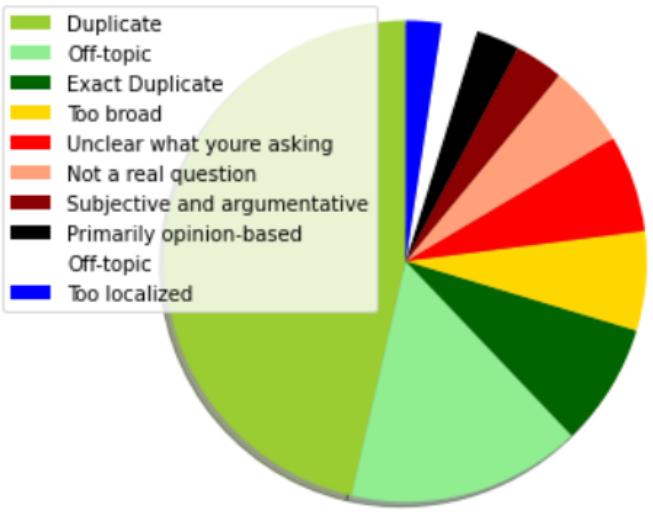

Fig. 3. shows the distribution of all five sub-categories of closed questions in our dataset.

topic, exact duplicate are the most dominant closed question categories and too localized is the less dominant category.

B. *Temporal Distribution Analysis*

By analyzing the presence of 'closed' questions on Stack Overflow over a 120-month time window between August 2008 to August 2012. Figure 4 depicts the ratio of 'closed' questions from 2010 to 2012 to total questions over this time period. Overall, the finding is an increasing trend of the percentage of 'closed' questions in each category. I found number of questions 'closed' over time has an upward curve. The most common categories of 'closed' questions over 120-months are Exact Duplicate and Not a Real Question. Both these categories dominate in presence over the others across.

C. *Effect of New Registered*

fig 5 shows a line graph shows number of closed questions asked by newly registered users over all closed questions asked in that month. The time range shown in the graph is in between 2008-08 to 2011-08.

D. *Community participation*

A question gets votes before getting closed. If a question in stack overflow gets five votes than the question gets automatically closed and if a moderator votes once the question also gets closed. fig 6 shows that more than 50 percent questions were closed in between 2008 to 2010 due to 5 votes and less than 10 percent for the other votes.

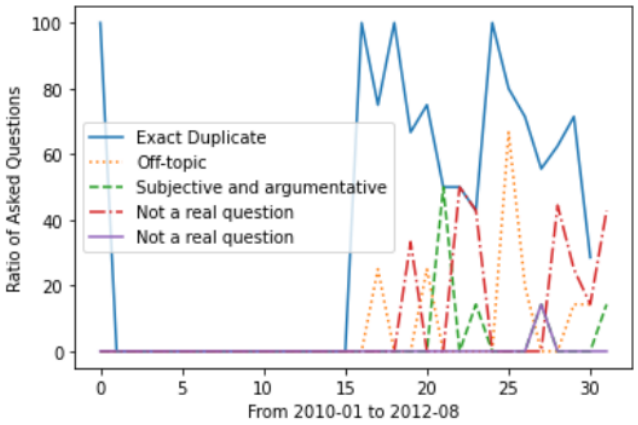

Fig. 4. shows the temporal distribution of all five sub-category.

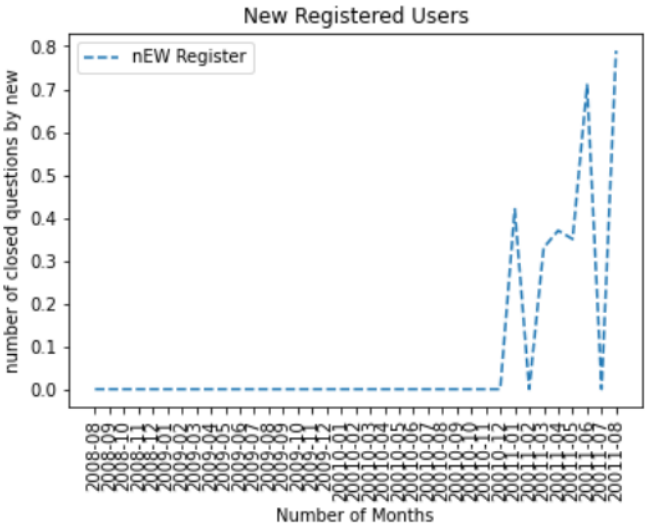

Fig. 5. shows the temporal distribution of all five sub-category.

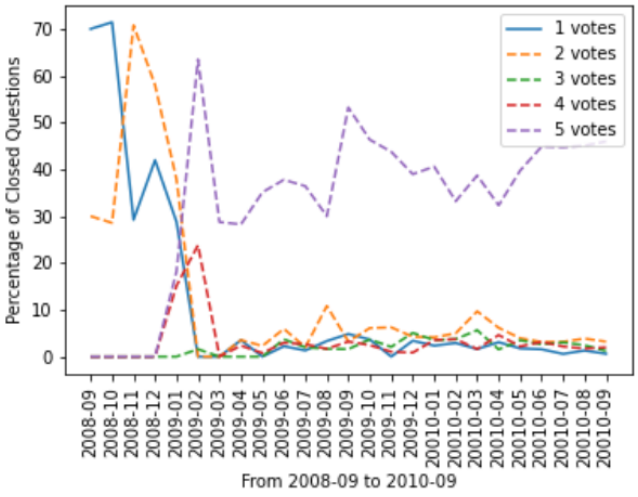

Fig. 6. shows the temporal distribution of 'close votes' in closed questions over a 24-month period from August 2008 to August 2010 We observe that a high percentage of questions are closed due to 2 votes

E. Topic analysis

A question asked in stack overflow has tags associated with it. In the finding the most used tag for closed question was JAVA. And CSS was at 20th place.

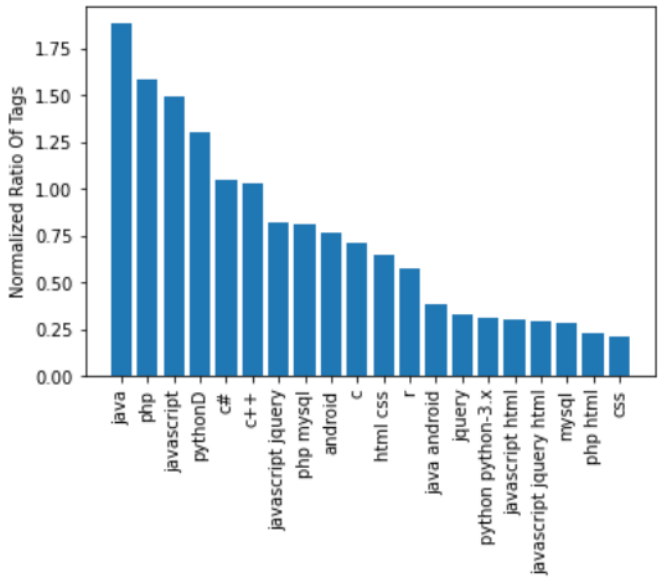

Fig. 7. shows the tags of closed questions on Stack Overflow with top 20 Normalized Tag Ratio

F. Content analysis

Here I characterized the content of 'closed' questions on Stack Overflow based on code snippet, tags, title and body. Every question doesn't contain code snippets. Only the valid questions should have snippets. In the analysis I found that even in the closed questions had code snippets. Figure 8 shows the percentage of code snippets found in each closed question category.

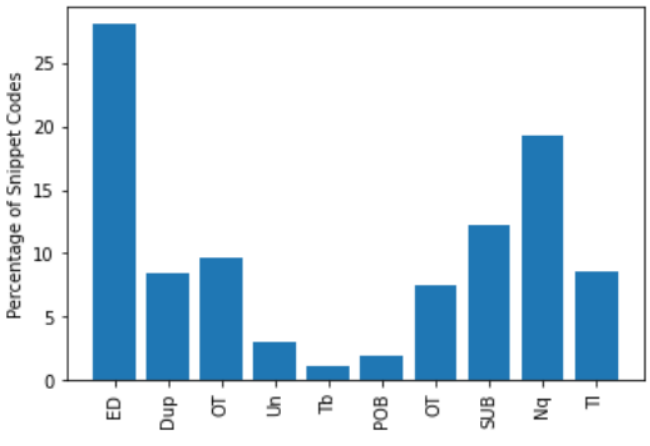

Fig. 8. Code Snippet for Closed Question categories

I found that Exact Duplicate and Not a real question category contains a large number of questions which have code snippets in them. The Exact Duplicate category by definition contains duplicate information to an existing question which may explain the high number. We see that the Too Broad contains the least amount of code snippets.

Figure 9 shows the unique tags in each closed question category. Most unique tags were found in the Exact Duplicate category and least was found in the Too Broad category.

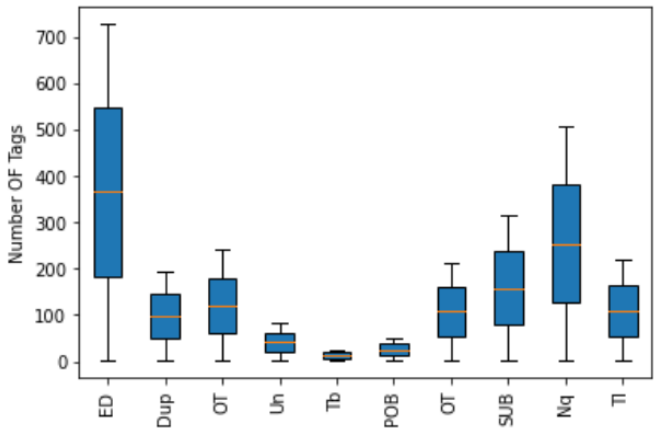

Fig. 9. Unique Tags

No Major differences were found in the Title length and Body. Both distributions has outliers and the medians are approximately similar in both of the categories. However, in the title length Too broad has the lowest median value and in body length not a real question and too localized has the lowest median. It indicates that questions belonging to this category are not a good fit to the Stack Overflow Question and Answer format even in terms of content. Figure 10 and 11 shows the character length distribution of question title and body in a box plot.

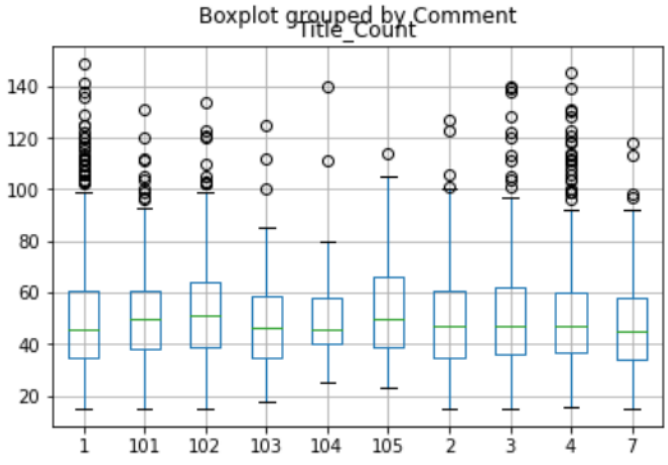

Fig. 10. Characters in Title length

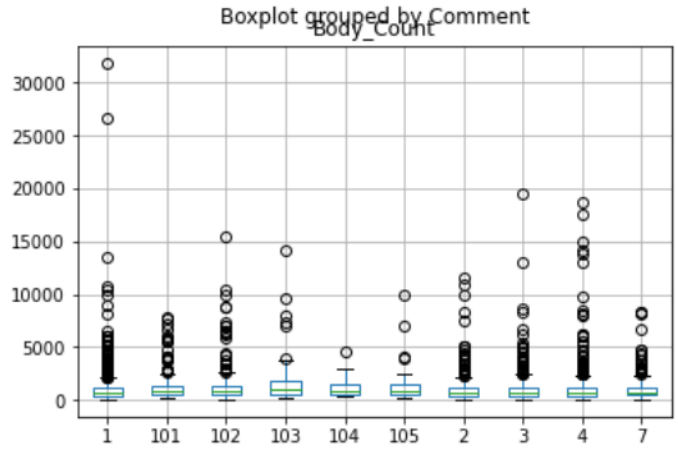

Fig. 11. Characters in body length

G. *Closure time*

By analyzing the time taken to 'close' questions on Stack Overflow. Figure 12 shows the closure time distribution of 'closed' question for every sub-category. The median closure times for Duplicate, Off Topic and Too Broad is 2.3le6, 2.2le6 and 2.1le6 minutes respectively. Most questions in these categories are quickly turned towards closure which may signify that their community value is relatively low than other categories. The Subjective has the highest median closure time. The reason for high closure time for the Subjective and Duplicate category could be because most questions (despite not being a good fit) invite discussion and opinions on broad programming related principles, guidelines, polls etc. Therefore, it takes time before these questions are answered in entirety and hence are left open for a longer time.

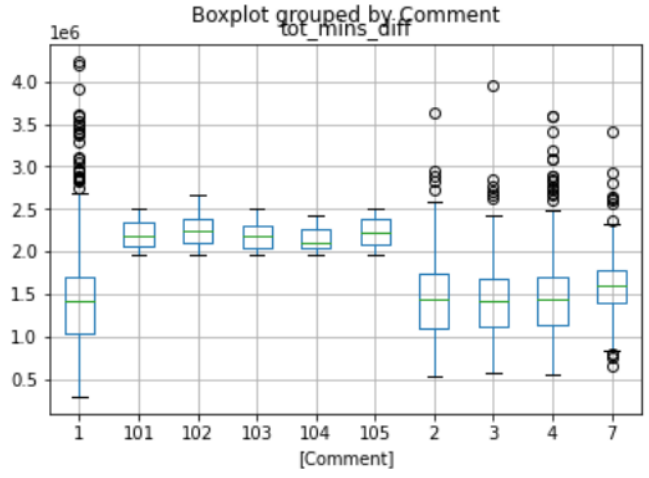

Fig. 12. shows the distribution of time taken to close questions for each category in minutes in the box and whisker plot

REFERENCES

[1] Denzil Correa, Ashish Sureka, Fit or Unfit : Analysis and Prediction of 'Closed Questions'on Stack Overflow.27 July 2013