QUANTITATIVE AND QUALITATIVE ANALYSIS OF AI AND ML PROJECTS ON GITHUB BY THE FIRST-TIME CONTRIBUTORS

Vivek AR and Karthikeyan P

Department of Information Technology, Thiagarajar College of Engineering, Madurai, Tamilnadu, India

ABSTRACT

The terms "machine learning" (ML) and "artificial intelligence" (AI) are widely used today. AI is working with many algorithms that include ML also. However, novice users use these two phrases individually. Analyzing and Understanding the significance and role of First Time Contributors in AI and ML projects helps to improvise the projects' content and provides them an opportunity to take up new projects. This work is presented with quantitative and qualitative analysis of AI and ML projects on GitHub. There are three research questions (RQ) prepared to support the analysis. The analysis is made by considering many parameters such as programming languages, forked repositories and commits.

KEYWORDS

Artificial Intelligence, Machine Learning, GitHub, Quantitative Analysis, Qualitative Analysis, Open-Source Projects, First-Time Contributors

1. INTRODUCTION

The Artificial Intelligence (AI) and Machine Learning (ML) based projects have seen a tremendous increase over the past five years and will have a significant role in the software industry in the upcoming years. AI and ML are closely related to Automation. According to yearly reports published by GitHub, there is a 43% improvement in proprietary development environments and 27% in open-source projects. So, a set of new developers contribute to these projects to improve their effectiveness and efficiency. Over millions of people use GitHub to discover, fork, and contribute to AI and ML projects. Also, these projects have found a new audience and set of possibilities in the development arena. This article focuses mainly on the first-time contributions made by the new developer community. Moderators can utilize the data used in this article to determine and call in new first-time developers. Adding on, it makes the job of more experienced users more straightforward and manageable. Tasks can be separated and given to first-timers so that there can be an efficient use of resources. Beginners are often provided with more manageable tasks by the moderators. That helps the first-timers adapt to new environments in the development phase. We used several python scripts for web scraping the data from GitHub. After the collection of the data, then we sorted it down using particular parameters and got a complete data set. The criteria used involves primarily similar to the contributions made by the first-timers. The article includes both Quantitative and Qualitative Analyses of the data collected. Eventually AI and ML projects are similar to one another also they include lots of algorithms like decision trees into them. But this work has been done with the classification between AI and ML projects with the AI and ML keywords. The following are the research questions (RQ) of this article:

RQ 1: What programming languages are preferred by first-timers in AI and ML projects?
RQ 2: What types of commits are made by first-time contributors in AI and ML projects?
RQ 3: What is the size of contribution made by a first-time contributor in AI and ML projects?

2. DATA COLLECTION

The data used in this article are first extracted from GitHub's website using several web scraping python scripts. That is done based on several criteria, including programming languages used, forked repositories, commits per repository, public repositories by the user, and the number of lines modified by the user.

2.1. Programming Languages

We have considered the first 5 to 6 languages used frequently in AI and ML projects, including Python, Notebook, JavaScript, Java, HTML, and C++. The remaining languages are considered under the 'Others' category. Notebook has been considered under this category because of the projects which implement them use a combination of other languages like Python, Scala combined with HTML, MATLAB combined with HTML, JavaScript. Also, the projects which are considered under each category are not duplicated as the projects are given a particular label for programming language by GitHub.

2.2. Forked Repositories

Using a similar approach to programming languages, we have extracted the forked repositories alone in this section.

2.3. Commits Per Repositories

In this criteria, we have shown types of commits and their corresponding number of modifications based on additions, deletions, file rename, changes, and other categories, which include Minor Bug Fixes and Typos.

2.4. User Repositories

Users with and without public repositories are considered. First-time contributors are considered as users with no Public Repository.

2.5. Number of Lines Modified

Modifications done by the user are classified based on the number of lines modified per commit from 1-5 Lines, 6-10 Lines, 11-100 Lines, and more than100 Lines.

3. RESULTS OF QUANTITATIVE ANALYSIS

3.1. AI Projects

Programming Languages - According to the data collected from 630690 projects, we observed that nearly 188961 projects (30%) were built with Python as language, followed by Jupyter Notebook and JavaScript with almost similar contributions about 123102 (20%) and 121190 (19%) respectively. Java and HTML have similar results contributing to about 68948 projects (11%) and 61307 projects (10%), respectively. With a minor percentage of contributions to C++,

there are about 31495 (5%) projects. In the Others Category, 35687 projects are contributed, making 5% of entire projects on AI from Table 1.

Forked Repositories - Out of 630690 projects on GitHub, only 403452 projects come under the category of forked repositories, i.e., 64% of the total repositories. They fairly spread across the type of programming languages as for Python with 121646 (64% of 188961) projects, Jupyter Notebook with 80860 (65% of 123102) projects, JavaScript with 79842 (65% of 121190) projects, Java with 42972 (62% of 68948) projects, HTML with 41179 (67% of 61307) projects, C++ with 21464 (68% of 31495) projects and other languages with 15489 (43% of 35687) projects from Table 2.

Table 1: Types of Programming Languages used in AI projects

Programming Language	**Repositories**	**Percentage**
Python	188961	30
Jupyter Notebook (Python, Scala…)	123102	20
JavaScript	121190	19
Java	68948	11
HTML	61307	10
C++	31495	5
C#,CSS,Ruby,TypeScript(Others)	35687	5

Table 2: Forked Repositories used in AI projects (relative to Table 1)

Programming Language	**Repositories**	**Percentage**
Python	121646	64
Jupyter Notebook (Python, Scala…)	80860	65
JavaScript	79842	65
Java	42972	62
HTML	41179	67
C++	21464	68
C#,CSS,Ruby,TypeScript(Others)	15489	43

Table 3: Commits in AI projects

Type Of Commit (per Repo)	Percentage
File Rename	26
Readme Changes	16
Deletions	14
Deletions (<10)	62
Deletions (11-100)	35
Deletions (>100)	3
Additions	24
Additions (<10)	51
Additions (11-100)	30
Additions (101-1000)	8
Additions (>1000)	11
Minor Bug Fixes, Typos (Others)	20

Commits per Repository - GitHub verified first 1000 repositories are considered in this section. We have classified the commits categories into File Rename with 26% of the changes; Readme Changes with 16%, Deletions in a file with 14%, Additions in a file with 24%, and Minor Bug Fixes, Typos in Others category with 20% of changes from Table 3.

User Repositories - Users with no public repositories include about 35% of the contributors, and the rest, 65%, contribute with at least one public repository to their name from Table 4.

The number of lines modified - We considered the first 1000 repositories that GitHub verifies. According to the data collected regarding AI projects, about 694 repositories in 1000 contribute to 1 to 5 Lines changed per commit. These account for about 69% of the total modifications made by the first-timers. In the other categories, 128 repositories changed 6-10 Lines and 109 repositories, with 11 to 100 Lines making it 13% and 11%, respectively. Changes with more than 100 Lines saw a significant drop to 69 repositories (i.e.) about 7% of the changes from Table 5.

Table 4: User data in AI projects

Types Of Users	Repositories	Percentage
No Public Repository	347	35
1-5 Public Repository	306	31
6-10 Public Repository	112	11
>10 Public Repository	235	23

Table 5: Number of Lines Changed in AI projects

Lines Modified per Commit	Repositories	Percentage
1-5 Lines	694	69
6-10 Lines	128	13
11 to 100 Lines	109	11
>100 Lines	69	7

3.2. ML Projects

Programming Languages - We have considered data collected out of 582811 projects. Users contributing to Jupyter Notebook projects are 305416 (52%), and Python projects 147379 (25%). The rest of the contributions belong to HTML, MATLAB, and the Others category with 48302 (8%), 34724 (5%), and 46990 (8%), respectively, from Table 6.

Table 6: Types of Programming Languages used in ML projects

Programming Language	**Repositories**	**Percentage**
Jupyter Notebook (Python, Scala…)	305416	52
Python	147379	25
HTML	48302	8
MATLAB	34724	5
R,JavaScript,Java,C++,C#,TeX(Others)	46990	8

Forked Repositories - According to the previous section data, only 59% of projects are considered in this section. Their distribution across the programming languages is also almost equal in percentage variation. The count of repositories with Jupyter Notebook is 182505, which is 60% of 305416 projects, Python having about 85410, which is 58% of 147379 projects, HTML with 26385 projects accounting for 55% of 48302 projects, MATLAB with 18773 projects making it 54% of 34724 projects and the other languages with 27935 projects about 59% of 46990 projects from Table 7.

Table 7: Forked Repositories used in ML projects (relative to Table 6)

Programming Language	**Repositories**	**Percentage**
Jupyter Notebook (Python, Scala…)	182505	60
Python	85410	58
HTML	26385	55
MATLAB	18773	54
R,JavaScript,Java,C++,C#,TeX(Others)	27935	59

Table 8: Commits in ML projects

Type Of Commit (per Repo)	Percentage
File Rename	34
Readme Changes	17
Deletions	12
Deletions (<10)	68
Deletions (11-100)	23
Deletions (>100)	9
Additions	19
Additions (<10)	56
Additions (11-100)	10
Additions (101-1000)	6
Additions (>1000)	20
Minor Bug Fixes, Typos (Others)	18

Commits per Repository - Considering only the first 1000 repositories that GitHub officially verifies, File Rename changes contribute to about 34% of changes, Readme Changes with about 17%, Deletions in a file of 12%, Additions in a file with 19%, and Minor Bug Fixes, Typos in Others category with 18% changes from Table 8.

User Repositories - About 54% of the users do not have a public repository, and the rest, 46%, do have a public repository in context to ML Projects from Table 9.

Table 9: User data in ML projects

Types Of Users	Repositories	Percentage
No Public Repository	539	54
1-5 Public Repository	156	16
6-10 Public Repository	103	10
>10 Public Repository	207	20

Table 10: Number of Lines Changed in ML projects

Lines Modified per Commit	Repositories	Percentage
1-5 Lines	648	65
6-10 Lines	154	15
11 to 100 Lines	114	11
>100 Lines	89	9

The number of lines modified - We also evaluated the first 1000 repositories verified by GitHub in the case of ML projects. About 648 repositories fell under the category of 1-5 Lines changed, making it 65% of the total 1000 repositories considered. For 6-10 Line changes and 11-100 Line changes, we have 154 repositories and 114 repositories contributing to 15% and 11%. With more than 100 Lines, 89 repositories made it 9% of the modifications from Table 10.

4. RESULTS OF QUALITATIVE ANALYSIS

4.1. AI Projects

4.1.1. Commits Per Repository

Deletions – contribute to about 14% of total changes done in a file. These are divided further in terms of frequency of deletions made per File as fewer than ten changes topped the list with 62% of the changes, 11-100 changes hold 35%, and the remaining 3% are made deletions with more than 100 lines. Additions – contributing 24% of the total number of changes made in a file. By counting the frequency of additions made per File, it can be observed that more than half of the additions belong to lesser than ten additions per File, 11-100 additions having 30% of share, 101-1000 additions with 8% of changes, and more than 1000 additions of 11% changes.

4.1.2. Public Repositories

No Public Repositories – From the collected data in 1000 users who contributed to the projects with 347 users contributing 35%. With Public Repositories – The remaining 65% of the users belong to this category. We have sub-categorized these repositories as 1-5 Public Repos, which involve 306 users (31%), 6-10 Public Repos containing 112 users (11%), and more than 10 Public Repos with 235 users (23%).

4.2. ML Projects

4.2.1. Commits Per Repository

Deletions – About 12% of file changes are deletions. We also sub-categorized them in terms of frequency of deletions made per File: 68% of deletions are less than 10 per File, 11-100 deletions of 23%, and the remaining 9% of changes come in more than 100 lines of deletions done by users. Additions – 19% of the commits involve additions. Of the 19% of them, 56% denote lesser than ten additions, 11-100 additions contributing about 10%, 6% of changes being 101-1000, 20% of additions coming under 1001-10000, and more than 10000 is 8% of additions.

4.2.2. Public Repositories

No Public Repositories – Of the 1000 users, 539 do not have a public repository. More than half the users have no public repository. With Public Repositories – About 156 users have 1-5 Public repositories, 103 users have 6-10 public repositories, and more than 10 Public repositories by 207 people.

5. CONCLUSIONS

Programming Languages – When we compared the results from AI and ML projects, we observed a similarity in the type of programming language used. Python and Jupyter Notebook contribute to more than 50% of projects on GitHub and adding JavaScript to the list it has been upgraded to more than 65% of the projects related to ML projects. We saw significant contributors choosing Jupyter Notebook and Python for AI projects, with more than 75% count. Forked Repositories – Contributors to AI and ML projects mostly choose the forked repositories. In this work, it can see that mostly 60% of the repositories are forked. As a result, most first-time contributors use forked repositories to make their contributions. Commits per Repository – First-time contributors are more interested in making Changes, Deletions, and Minor Bug Fixes and

Typos, as we can observe from the data presented in this article. Also, additions and file rename show minor variations, choosing first-time contributors on a smaller scale than the AI and ML projects. Public Repositories – Mostly, half of the users do not have a public repository concerning ML-related projects, but when it comes to AI projects, only about 35% of users are first-timers. We observed this change in the number of first-time contributors compared to that of AI projects. The number of Lines changes In comparison, we observed that about 80% of the commits were fewer than ten lines changes per File, which shows a higher significance of first-time contributors.

According to the analysis made in this work, for RQ 1: Python and Jupyter Notebook are the most favoured programming languages among the others used by first-timers for AI and ML projects. For RQ 2: Most of the commits made by the first-time contributors include minor changes to files, Deletions and Minor Bug Fixes and Typos. For RQ 3: Based on these observations, the extent of commits made by first-timers is not the largest in terms of size but are interested in making smaller contributions to open-source projects. In future, the work will be extended with the analysis by considering various subset algorithms under AI and ML. It will also be extended with other public repositories data set.

ACKNOWLEDGEMENTS

The authors would like to thank the Institution and Anonymous Reviewers.

REFERENCES

[1] V. Subramanian, I. Rehman, M. Nagappan and R. Kula, (2022) "Analyzing First Contributions on GitHub: What Do Newcomers Do?" in IEEE Software, vol. 39, no. 01, pp. 93-101, 2022.

[2] Riehle, D. (2015). The Five Stages of Open Source Volunteering. In: Li, W., Huhns, M., Tsai, WT., Wu, W. (eds) Crowdsourcing. Progress in IS. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-47011-4_2

[3] The State of the October - GitHub. https://octoverse.github .com/ (accessed Aug 27, 2022)

AUTHORS

Vivek AR is currently a sophomore pursuing Bachelor of Technology in Information Technology at Thiagarajar College of Engineering Madurai. He has completed HSC in the year 2021with 95.8% score and SSLC in the year 2019 with 93% score at St.John's English School and Junior College Besant Nagar Chennai. His research interests include website analysis, machine learning and software development.

Karthikeyan P is currently working as an Associate Professor in Thiagarajar College of Engineering, Madurai. He completed the Ph.D. programme in Information and Communication Engineering under Anna University, Chennai, Tamilnadu, India in the year 2015. He has 16 years of teaching and 6 years of research experience. He published many papers in refereed international journals, conferences and book chapters. He also published 4 Indian patents and filed 1 Indian patent. He is a reviewer in various international journals like IEEE Transactions on Cybernetics, Education, IEEE Access, etc. He served as a guest editor in IJMLO (Inderscience). His research interests include evolutionary algorithms, ad hoc networks, educational technology and machine learning.