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

K. Sindhu [1],Akshith Nag [2], Asra Anjum [3]

[1]Assistant Professor, Swarna Bharathi Institute of Science & Technology, India, E-mail: karlapudisindhu516@gmail.com.
[2]Assistant Professor, Swarna Bharathi Institute of Science & Technology, India, E-mail: Nag.686@gmail.com.
[3]Assistant Professor, Swarna Bharathi Institute of Science & Technology, India, E-mail: asraanju@gmail.com.

ABSTRACT-

The words "artificial intelligence" (AI) and "machine learning" (ML) have entered common use. AI is using a wide variety of algorithms, including ML. Nevertheless, inexperienced users often employ each of these expressions alone. Improving the substance of AI and ML projects and giving first-time contributors a chance to take on new challenges may be achieved via analyzing and understanding their relevance and function. Analyses of ML and AI projects hosted on GitHub, both quantitative and qualitative, constitute the bulk of this work. In order to back up the analysis, three research questions (RQ) have been generated. A number of factors, including programming languages, forked repositories, and commits, are taken into account in the study. Open-Source Projects, First-Time Contributors, Quantitative and Qualitative Analysis, Machine Learning, GitHub

INTRODUCTION

There has been a meteoric rise in the number of projects centered on AI and ML in the last five years, and these initiatives will continue to play a pivotal role in the software industry going forward. Automation is a close relative of AI and ML. Private development environments see a 43% improvement while open-source projects see a 27% improvement, according to GitHub's annual reports. In order to make these projects better and more efficient, a group of fresh developers are contributing to them. Discovering, forking, and contributing to AI and ML projects is a common practice among the millions of users that utilize GitHub. The development arena has also opened up new options and an audience for these ventures. The new developer community's first contributions are the primary topic of this article. Using the information gathered from this article, moderators may identify and invite new first-time developers. And it simplifies and manages the work of more advanced users as well. To make better use of resources, tasks might be divided up and assigned to people who are new to the process. It is common practice for moderators to provide easier duties to beginners. In the development phase, it aids first-timers in adapting to new settings. In order to get the data from GitHub, we employed a number of programs written in Python. Once the data was collected, it was sorted using certain characteristics to provide a comprehensive set of records. The contributions provided by the first-timers are mostly comparable to the criteria that are applied. The data presented in the paper have been analyzed using both quantitative and qualitative methods. At their core, AI and ML projects are quite similar, and both use a plethora of algorithms, such as decision trees. However, this effort has been completed by categorizing AI and ML projects using the relevant keywords. The article's research questions (RQ) are as follows:

DATA COLLECTION

Using a variety of web scraping python scripts, the data used in this study is first retrieved from the GitHub website. Several factors are taken into consideration for this, such as the languages utilized, the amount of forked repositories, the number of commits per repository, the user's public repositories, and the number of lines updated.

2.1.1 Languages Used for Programming Python, Notebook, JavaScript, Java, HTML, and C++ are the top five or six languages used in AI and ML applications. The 'Others' category includes all the other languages. The projects that employ Notebook fall into this category since they use a mix of languages such as Python, Scala, MATLAB, and JavaScript. Because GitHubassigns each project a unique title based on the programming language it uses, there is no duplication of projects that fall under each category.

2.2. Splitting Up Files In this part, we have isolated the forked repositories using a method that is comparable to programming languages.

2.3. One repository at a time commits Using this criterion, we have illustrated the different kinds of commits and the amount of changes for each, broken

down by categories such as typos, minor bug fixes, file renames, and additions/deletions Section
2.4: User Files We take into account users who have and do not have public repositories. Users without a Public Repository are categorized as first-tie contributors.
2.5. Total Lines Changed You may sort user modifications into four categories: 1–5 lines, 6–10 lines, 11–100 lines, and more than 100 lines per commit.

RESULTS OF QUANTITATIVE ANALYSIS

Languages Used for Coding - Out of 630690 projects, 188961 (30%) were constructed using Python. Jupyter Notebook and JavaScript came in second and third, with roughly identical contributions of around 123102 (20%) and 121190 (19%), respectively. Approximately 68948 projects (or 11% of all projects) use Java, whereas 61307 projects (or 10% of all projects) use HTML. Approximately 31495 projects (or 5% of the total) make contributions to C++, according to Computer Science & Engineering: An International Journal (CSEIJ), Vol 12, No 6, December 2022. Table 1 shows that 35687 projects, or 5% of all AI initiatives, fall under the Others Category.

Among the 630690 projects hosted on GitHub, 403452 (or 64% of the total) are considered to be forked repositories. From Table 2, we can see that the languages used for these projects are fairly diverse: 121646 (64% of 188961) were Python projects, 80860 (65% of 123102) were Jupyter Notebook projects, 42972 (62% of 68948) were Java projects, 41179 (67% of 61307) were HTML projects, 21464 (68% of 31495) were C++ projects, and 15489 (43% of 35687) were other languages.

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

RESULTS OF QUALITATIVE ANALYSIS

AI Initiatives

About fourteen percent of all file modifications occur as a result of deletions, which constitute 4.1.1. Commits Per Repository. Further categorization is provided by the frequency of deletions per file: 62% of modifications had ten lines or less, 35% had eleven to one hundred lines, and 3% had one hundred or more lines deleted. Adding new content to an existing file accounts for a quarter of almodifications. If we look at the frequency of additions per file, we can see that over half of the additions are in the 10–100 additions category, with 30% of the total. The 101–1000 additions category accounts for 8% of the modifications, while the more than 1000 additions category accounts for 11%.

4.2 ML Projects

4.2.1. Deletions Per Repository Commit

Roughly 12% of all file modifications are deletions. We then divided them according to the frequency of deletions per file: 68% of deletions are under 10 lines per file, 23% are between 11 and 100 lines, and 9% are more than 100 lines of deletions. Component additions account for 19% of all contributions. Out of the 19%, 56% point to additions of fewer than ten, 10% to additions of 11–100, 6% to changes of 101–1000, 20% to additions of 1001–10,000, and 8% to additions of more than 10,000 Section.

4.2.2: Public Data Stores

Private Data Stores Only - There are 539 users out of 1000 without a public repository. Half or more of the users do not have a public repository. In terms of public repositories, 156 users have between one and five, 103 have between six and ten, and 207 have ten or more.

CONCLUSION

We found that the outputs of AI and ML projects employed comparable types of programming languages when we compared them. More over half of all GitHub projects use Python and Jupyter Notebook; when you include JavaScript, that number jumps to almost 65% of all ML-related projects. For their artificial intelligence projects, over 75% of our contributors opted for Jupyter Notebook and Python. Branched Databases - The forked repositories are often chosen by contributors to AI and ML projects. The majority of the repositories, over 60%, are forked according to this study. Consequently, most people who contribute for the first time do so via forked repositories. The statistics offered in this post clearly shows that there are typos. Additionally, there are subtle differences in the contributions and file renaming processes, with fewer first-time contributors being selected compared to the AI and ML projects. Repositories Open to the Public — While 35% of users are new to AI projects, 50% of users do not have a public repository for ML-related projects. We saw this shift in the proportion of first-time donors when comparing AI initiatives to others. Line count fluctuates A greater proportion of first-time contributors were involved in the 80% of commits that were less than 10 lines of modifications per file. This study's analysis answers the first research question (RQ1) on the languages most often used by beginners for artificial intelligence (AI) and machine learning (ML) projects: Python and Jupyter Notebook. In response to RQ 2, the majority of first-time contributors' submissions consist of small file deletions, edits, and fixes for bugs and typos. Regarding RQ 3, these findings suggest that first-timers are engaged in providing modest contributions to open-source projects, rather than the greatest ones in terms of size. The study will be expanded in the future to include other subset methods under AI and ML. Additional public repository data sets will be added to it as well. The most common types of commits made by first-time contributors to a repository are changes, deletions, and minor bug fixes.

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)