Mining Software Repositories
Rapport de la Partie 2 : Implémentation et Analyse du Dataset AIDev

Nihed Jamaoui

1 Introduction

Cette seconde partie du travail vise à implémenter une analyse reproductible du dataset **AIDev**, dans le cadre d'une étude de *Mining Software Repositories (MSR)*. L'objectif principal est de charger, nettoyer, analyser et interpréter les données afin de répondre aux questions de recherche définies lors de la première partie du projet.

Ce rapport synthétise les étapes techniques réalisées, les résultats obtenus ainsi que les limites rencontrées.

2 Rappel du Travail de la Partie 1

Dans la première partie, nous avons :

- étudié la structure du dataset AIDev,
- défini un objectif de recherche clair,
- formulé trois questions de recherche (RQ1, RQ2, RQ3),
- identifié des métriques simples et mesurables.

Ces éléments constituent la base conceptuelle du travail implémenté dans cette seconde partie.

3 Implémentation : Pipeline Reproductible

L'implémentation a été réalisée à l'aide d'un **notebook Python**, garantissant la reproductibilité de l'analyse.

Les étapes principales du pipeline sont les suivantes :

1. Chargement du dataset AIDev depuis Hugging Face
2. Nettoyage des données (suppression des entrées incomplètes)
3. Définition d'un subset pertinent basé sur :
    - la présence d'un agent identifié,

- des pull requests fermées,
- des dates de création et de fermeture valides

4. Calcul des métriques définies
5. Génération de visualisations

Le subset final contient **859 927 pull requests** décrites par **14 attributs**.

4 Analyse et Résultats

L'analyse repose sur trois axes principaux :

4.1 Contribution des agents IA (RQ3)

L'analyse montre que la totalité des pull requests du dataset AIDev est générée par des agents IA. La contribution est fortement dominée par certains agents, notamment *OpenAI Codex*, suivi de *GitHub Copilot* et d'autres agents.

4.2 Devenir des pull requests IA (RQ1)

Les résultats indiquent :

- un taux de fusion élevé (environ 92%),
- une proportion limitée de pull requests fermées sans fusion.

Cette analyse reste partielle en raison de l'absence de pull requests humaines dans le dataset.

4.3 Temps de traitement des pull requests (RQ2)

Le temps moyen de traitement des pull requests générées par des agents IA est faible (environ 0.33 jour), ce qui suggère une validation rapide des contributions automatisées.

5 Discussion et Limites

Une observation importante issue de l'analyse est que le dataset AIDev ne contient aucune pull request générée par des développeurs humains. Par conséquent :

- les comparaisons IA vs humain (RQ1 et RQ2) ne peuvent être réalisées que de manière partielle,
- les résultats doivent être interprétés comme une caractérisation interne des contributions IA.

D'autres limites concernent :

- l'absence de métriques de qualité du code,
- la domination de certains agents pouvant biaiser les résultats globaux,
- la généralisation limitée à d'autres projets ou outils IA.

6 Lien avec le Papier Scientifique

Les résultats de cette implémentation ont été synthétisés et approfondis dans un papier scientifique de type conférence intitulé :

> *Mining Software Repositories: An Exploratory Study of AI-Generated Pull Requests Using the AIDev Dataset*

Ce papier adopte une structure académique complète (Introduction, Study Design, Analysis, Discussion, Threats to Validity, Conclusion) et constitue la synthèse finale du travail réalisé.

7 Conclusion

Cette seconde partie du projet a permis de mettre en œuvre une analyse MSR complète et reproductible sur le dataset AIDev. Malgré certaines limitations liées à la nature du dataset, le travail réalisé met en évidence le potentiel des agents de programmation IA dans les processus de développement logiciel et ouvre la voie à des études futures intégrant des contributions humaines.