IN PRESS

CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain

Mapping technological diversification in open-source software: A Transformer-based approach for analysing knowledge reuse in GitHub

Leonardo Mazzoni[1], Fabio Pinelli[2], Emilio Incerto[2], Mauro Lombardi[3], Mariano Mastrogiorgio[4]

[1]European University Institute, Robert Schuman Centre - Centre for a Digital Society, Italy, [2]IMT School for Advanced Studies Lucca, Computer Science, Italy, [3]University of Florence, Economics and Management, Italy, [4]IE University Business School, Strategy, Spain.

How to cite: Mazzoni, L.; Pinelli, F.; Incerto, E.; Lombardi, M.; Mastrogiorgio, M. (2026). Mapping technological diversification in open-source software: A Transformer-based approach for analysing knowledge reuse in GitHub. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA2026.2026.21904

Abstract

This paper presents a scalable methodological pipeline for constructing and analysing large-scale software reuse networks from GitHub data. Studying how software components are repurposed across projects and technological domains is of growing interest for innovation scholars, yet existing approaches lack scalable methods to capture technological proximity between digital artefacts. We address this gap by introducing a multi-stage pipeline that combines (i) dependency-based network extraction from GitHub's API, (ii) zero-shot topic extraction via the Gecko transformer architecture, (iii) embedding-based representation of repositories in a continuous semantic space, and (iv) hierarchical agglomerative clustering to identify technological domains. The pipeline processes approximately 96,000 dependency–dependent pairs across more than 45,000 repositories, generating a continuous measure of technological distance based on cosine similarity between repository-level embedding vectors. The approach offers a replicable framework for mapping diversification dynamics at the micro level in open-source software ecosystems, providing new research avenues for text-based analyses of technologically intensive industries

Keywords: *Technological Diversification; Innovation; Open-Source Software; GitHub; Large Language Models; Transformers*

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València