**Automatic detection of privacy violations in GitHub projects**

SE-2-3

By: Tomer Ben Shimol; tomerbe3@ac.sce.ac.il  
 Eliyahu Amuyev; eliam2@ac.sce.ac.il

Advisors: Dr. Marina Litvak 1

Jenny Guber2

1Shamoon College of Engineering, Beer-Sheva

2University of Haifa

Privacy violations occur when personal information is accessed or used without consent, and detecting them is important to prevent harm and protect privacy. In this project, we trained machine-learning models that can accurately classify issues related to privacy violations. We collected and annotated a new dataset of issues reported in GitHub projects. Then, we experimented with different text representation and classification models and reported their evaluation results on this dataset. Our findings can assist GitHub users in managing privacy issues more effectively. Improved classification could enhance trust and protect sensitive information, benefitting social media and software development collaboration platforms.

**Keywords:** Privacy, Privacy violations, classification, issues.