

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: May 25, 2024 at 3:15 PM

Speaker Information

Zahin Wahab (Std No. 0421052013) is a part-time M.Sc. student in the Department of CSE, BUET. Moreover, she is currently working as a lecturer in the Department of CSE at BRAC University. She completed her B.Sc. with honors from CSE, BUET in 2021. Her research interests lie in the field of programming languages and software engineering. She is currently doing her postgraduate thesis under the supervision of Prof. Dr. Rifat Shahriyar. She will be speaking about her ongoing research in this talk.

Secret Breach Prevention in Software Issue Reports

In the digital age, the exposure of sensitive information poses a significant threat to security. Leveraging the ubiquitous nature of code-sharing platforms like GitHub and BitBucket, developers often accidentally disclose credentials and API keys, granting unauthorized access to critical systems. Despite the availability of tools for detecting such breaches in source code, detecting secret breaches in software issue reports remains largely unexplored. This thesis presents a novel technique for secret breach detection in software issue reports using a combination of language models and state-of-the-art regular expressions. We highlight the challenges posed by noise, such as log files, URLs, commit IDs, stack traces, and dummy passwords, which complicate the detection process. By employing relevant pre-processing techniques and leveraging the capabilities of advanced language models, we aim to mitigate potential breaches effectively. Drawing insights from existing research on secret detection tools and methodologies, we propose an approach combining the strengths of state-of-the-art regexes with the contextual understanding of language models. Our method aims to reduce false positives and improve the accuracy of secret breach detection in software issue reports. We have also developed a secret breach mitigator tool for GitHub, which will warn the user if there is any secret in the posted issue report. By addressing this critical gap in contemporary research, our work aims at enhancing the overall security posture of software development practices. We have curated a benchmark dataset of 25000 instances with only 437 true positives. Although the data is highly skewed, our model performs well with a 0.6347 F1-score, whereas state-of-the-art regular expression hardly manages to get a 0.0341 F1-Score with very poor precision.