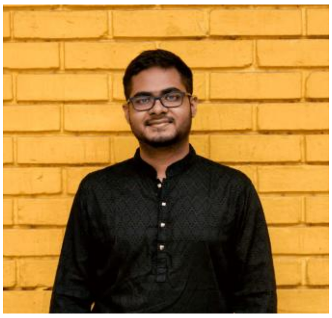
Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: March 07, 2026 at 1:30 PM

Speaker Information



Md Awsaf Alam Anindya (Std No. 0421052071) is a full time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies from BUET in 2021. His research interest lies in the fields of Software Engineering. He is currently doing his postgraduate thesis under the supervision of Dr. Anindya Iqbal. He will be speaking about his ongoing research in this talk.

Promoting Inclusive Developer Communication through Real-Time Toxicity Filtering

Toxic interactions during code reviews can undermine teamwork and hinder productivity in software engineering (SE) teams. While prior studies explore toxicity detection and empirical investigation, they lack real-time detoxification tools to support the SE community. This thesis presents ToxiShield, a comprehensive framework for promoting inclusive developer communication through real-time toxicity filtering in GitHub pull request reviews.

ToxiShield is built using three modules: (i) Module 1 – Toxicity Filter, which identifies whether a text is toxic using a fine-tuned BERT model; (ii) Module 2 – Communication Coach, which facilitates just-in-time fine-grained toxicity categorization with explanations powered by large language model; and (iii) Module 3 – The Reframer, which generates a revised, constructive alternative of a toxic text while preserving technical intent. The system is implemented as a browser extension that operates proactively, identifying and mitigating toxicity as developers write comments, rather than reactively flagging content after publication.

Our BERT-based binary detection model, trained on 38,761 code review samples, achieves 98% accuracy and an F1-score of 97%. The multiclass toxicity classification component provides detailed reasoning for toxicity categorization across 12 toxicity categories, with Claude 3.5 Sonnet achieving 42% macro F1-score and 39% macro MCC in multiclass toxicity classification with detailed reasoning. For detoxification, we fine-tune five large language models on 10,120 code review comments, with the fine-tuned Llama 3.2 model achieving 95.27% style accuracy, 97.03% fluency, 67.07% content preservation, and an 84% J-score.

The system's effectiveness is validated through comprehensive evaluation, including automated metrics and human evaluation studies. A user study with 10 software developers using the Technology Acceptance Model confirms the tool's perceived usefulness and ease of adoption, with high satisfaction scores across all dimensions. ToxiShield represents the first real-time detoxification tool specifically designed for software engineering environments, setting a benchmark for advancing constructive communication and fostering inclusivity in open-source communities.