INFORMATION EXTRACTION FROM SCIENTIFIC LITERATURE ON EMS MUTAGENESIS USING TEXT MINING AND BERT MODEL

Supervised by Ery Setyawan Jullev Atmadji S.Kom, M.Cs.

Maulana Malik Ibrahim

Study Program of Informatics Engineering

Majoring in Information Technology

ABSTRACT

Chemical mutagenesis using Ethyl Methane Sulfonate (EMS) is one of the primary methods in plant breeding programs for inducing controlled genetic diversity. Information regarding EMS treatment parameters, including dosage, target organism, mutated genes, and phenotype, is scattered across thousands of unstructured scientific publications, making it difficult for plant breeders to efficiently access the accumulated knowledge. This study develops an automated information extraction system based on the Bidirectional Encoder Representations from Transformers (BERT) model, specifically BioBERT, fine-tuned for the Named Entity Recognition (NER) task on the EMS mutagenesis domain. To address the data scarcity challenge inherent to the molecular biology domain, this study integrates an Active Learning approach with a Teacher-Student Model pseudo-labeling mechanism. Through eight Active Learning cycles, the training data expanded automatically from 1,252 manually annotated sentences to 12,037 high-quality sentences. The resulting model achieved an overall F1-Score of ***93.13%*** *(Precision 92.54%, Recall 93.72%) under the Strict Exact Match evaluation. Per-entity performance yielded F1-Scores of 95% for the GENE class, 93% for ORGANISM, 91% for PHENOTYPE, and 90% for DOSAGE. The system was implemented using a Microservices architecture integrating Laravel as the frontend and Flask API as the artificial intelligence backend, augmented by a Continuous Learning (Data Flywheel) mechanism that enables the system to improve continuously with increased usage. The results demonstrate that a BioBERT-based text mining approach provides a reliable and effective tool for comprehensively synthesizing EMS mutagenesis knowledge from scientific literature without requiring manual reading.*

Keywords: *Active Learning*, BioBERT, *Ethyl Methane Sulfonate* (EMS), *Information Extraction*, Mutagenesis, *Named Entity Recognition*, *Text Mining*.