UNIVERSITAS
GADJAH MADA
**Deteksi Code Smell Menggunakan Kombinasi Metrik Struktural dan Representasi Semantik Kode pada
Dataset yang Relevan dengan Industri**
Azhar Bagaskara, Arif Nurwidyantoro, M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

DETEKSI *CODE SMELL* MENGGUNAKAN KOMBINASI METRIK STRUKTURAL DAN REPRESENTASI SEMANTIK KODE PADA DATASET YANG RELEVAN DENGAN INDUSTRI

Oleh

AZHAR BAGASKARA
22/502652/PA/21573

Keberadaan *code smell* dapat mempersulit pemeliharaan perangkat lunak. Deteksi *code smell* secara otomatis telah berkembang ke pendekatan pembelajaran mesin (*machine learning*), namun sebagian besar penelitian terdahulu hanya berfokus pada salah satu jenis fitur saja (metrik struktural atau representasi semantik). Penelitian ini mengusulkan model deteksi berbasis pembelajaran mesin dengan menggabungkan metrik struktural (metrik CODEBEAT dan PMD) dan representasi semantik kode (vektor *embedding* code2vec, CuBERT, dan CodeBERT) pada dataset MLCQ yang relevan dengan industri. Klasifikasi dilakukan menggunakan algoritma *random forest*, XGBoost, dan SVM dengan label hasil *majority voting*. Hasil eksperimen menunjukkan bahwa penggabungan fitur mampu meningkatkan performa deteksi, dengan tingkat peningkatan yang bervariasi tergantung karakteristik unik *code smell* (cukup terlihat pada *data class*, namun marginal pada *long method* dan *feature envy*). Performa terbaik dicapai pada deteksi *data class* dengan *F1-score* $0,67 \pm 0,04$ dan MCC $0,62 \pm 0,05$ menggunakan *random forest* dan *embedding* CuBERT atau CodeBERT. Sementara itu, deteksi *blob* mendapatkan *F1-score* $0,53 \pm 0,04$ menggunakan SVM dan CodeBERT, deteksi *long method* mencatat *F1-score* $0,77 \pm 0,04$ menggunakan XGBoost dan code2vec, serta deteksi *feature envy* mencapai *F1-score* $0,30 \pm 0,07$ menggunakan *random forest* dan CuBERT. Hasil ini menunjukkan potensi kombinasi fitur dalam meningkatkan akurasi deteksi *code smell* di industri.

Kata-kata kunci: Deteksi *Code Smell*, Metrik Struktural, Representasi Semantik Kode, Pembelajaran Mesin, Dataset MLCQ, Random Forest, XGBoost, SVM, code2vec, CuBERT, CodeBERT

UNIVERSITAS
GADJAH MADA
**Deteksi Code Smell Menggunakan Kombinasi Metrik Struktural dan Representasi Semantik Kode pada
Dataset yang Relevan dengan Industri**
Azhar Bagaskara, Arif Nurwidyantoro, M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

CODE SMELL DETECTION USING A COMBINATION OF STRUCTURAL METRICS AND SEMANTIC CODE REPRESENTATIONS ON AN INDUSTRY-RELEVANT DATASET

By

AZHAR BAGASKARA
22/502652/PA/21573

The presence of code smells complicates software maintenance. While automated code smell detection has shifted towards machine learning, previous studies typically focus on either structural metrics or semantic representations, limiting comprehensive understanding. This study proposes a machine learning-based detection model combining structural metrics (CODEBEAT and PMD metrics) and semantic representations (code2vec, CuBERT, and CodeBERT embedding vectors) on the industry-relevant MLCQ dataset. Models are classified using random forest, XGBoost, and SVM, with labels derived from majority voting. The experimental results show that combining features improves detection performance, though the improvement varies based on the unique characteristics of each code smell (noticeable for data class, but marginal for long method and feature envy). The best performance is achieved in data class detection with an F1-score of 0.67 ± 0.04 and MCC of 0.62 ± 0.05 using random forest with CuBERT or CodeBERT embeddings. Meanwhile, blob detection achieves an F1-score of 0.53 ± 0.04 using SVM with CodeBERT, long method detection obtains an F1-score of 0.77 ± 0.04 using XGBoost with code2vec, and feature envy detection reaches an F1-score of 0.30 ± 0.07 using random forest with CuBERT. These findings demonstrate the potential of combined feature models in improving code smell detection to support software maintenance in the industry.

Keywords: Code Smell Detection, Structural Metrics, Semantic Representation of Code, Machine Learning, MLCQ Dataset, Random Forest, XGBoost, SVM, code2vec, CuBERT, CodeBERT