UNIVERSITAS
GADJAH MADA
Deteksi Duplikasi Laporan Bug Berbasis Klasifikasi Menggunakan Arsitektur Single-Instance dan Two-Stage Pairwise
Muhammad Hariish Hafiiz, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR ISI

HALAMAN JUDUL ii

HALAMAN PENGESAHAN iii

PERNYATAAN iv

KATA PENGANTAR v

DAFTAR ISI vi

DAFTAR TABEL xi

DAFTAR GAMBAR xiii

DAFTAR ALGORITMA xv

DAFTAR KODE xvi

INTISARI xvii

ABSTRACT xviii

I PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Rumusan Masalah 4

1.3 Batasan Masalah 4

1.4 Tujuan Penelitian 5

1.5 Manfaat Penelitian 6

1.6 Sistematika Penulisan 6

II TINJAUAN PUSTAKA 8

2.1 Pendekatan Berbasis *Information Retrieval* 8

2.2 *Pendekatan Berbasis* Machine Learning *dan* Deep Learning 11

III LANDASAN TEORI 16

3.1 Bug Perangkat Lunak 16

UNIVERSITAS
GADJAH MADA
Deteksi Duplikasi Laporan Bug Berbasis Klasifikasi Menggunakan Arsitektur Single-Instance dan Two-Stage Pairwise
Muhammad Hariish Hafiiz, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

3.2 Deteksi *Bug Report* Duplikat (DBRD) 16
3.3 *Natural Language Processing* (NLP) 17
3.4 Dataset Proyek Eclipse 17
3.5 Pra-pemrosesan Teks (*Text Preprocessing*) 17
3.6 Teknik Representasi Fitur (*Feature Engineering*) 18
3.6.1 Representasi Leksikal (Berbasis IR) 19
3.6.2 Representasi Semantik Kontekstual 21
3.6.3 Representasi Topik (*Topic Modeling*) 26
3.6.4 Pengukuran Kemiripan: *Cosine Similarity* 27
3.7 Algoritma Klasifikasi 28
3.7.1 *Logistic Regression* 28
3.7.2 *Random Forest* 29
3.7.3 *Extreme Gradient Boosting* (XGBoost) 29
3.7.4 *Multi-Layer Perceptron* (MLP) 30

IV METODE DAN RANCANGAN PENELITIAN 31
4.1 Deskripsi Umum Penelitian 31
4.2 Analisis Permasalahan 33
4.2.1 Inefisiensi Komputasi pada Arsitektur *Pairwise* 34
4.2.2 *Trade-off* Efisiensi dan Akurasi melalui Arsitektur Alternatif 34
4.2.3 Kesenjangan Representasi Fitur Teks dan Atribut Laporan 35
4.3 Dataset dan Mekanisme Pelabelan 35
4.3.1 Sumber dan Atribut Data 36
4.3.2 Definisi Label dan *Ground Truth* Arsitektur *Single-Instance* 37
4.3.3 Definisi Label dan *Ground Truth* Arsitektur *Two-Stage Pairwise* 38
4.3.4 Korpus Retrieval 39
4.3.5 Pembagian Data 40
4.4 Rancangan Algoritma Pelatihan 40
4.4.1 Pra-pemrosesan Teks 41
4.4.2 Ekstraksi dan Fusi Fitur 42
4.4.3 Arsitektur *Single-Instance* 46
4.4.4 Arsitektur *Two-Stage Pairwise — Stage* 1 (*Candidate Retrieval*) 48
4.4.5 Arsitektur *Two-Stage Pairwise — Stage* 2 (*Pairwise Classifier*) 48
4.4.6 *Baseline Pairwise* (Replikasi Ben Messaoud) 53
4.5 Penyetelan *Hyperparameter* 55

UNIVERSITAS
GADJAH MADA
Deteksi Duplikasi Laporan Bug Berbasis Klasifikasi Menggunakan Arsitektur Single-Instance dan Two-Stage Pairwise
Muhammad Hariish Hafiiz, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

4.5.1 Penyetelan *Hyperparameter Classifier* (SI dan *Stage* 2) 55
4.5.2 *Hyperparameter Fine-tuning* SBERT (*Stage* 1) 56
4.5.3 *Hyperparameter* Pelatihan *Baseline* BERT 56
4.6 Metrik Evaluasi 57
4.6.1 Metrik Kualitas Deteksi 57
4.6.2 Metrik Efisiensi 59
4.7 Skenario Eksperimen Perbandingan 60
4.7.1 Eksperimen 1: Kualitas *Pairwise Classifier* terhadap *Baseline* 60
4.7.2 Eksperimen 2: Efektivitas Rangkaian *Two-Stage* 60
4.7.3 Eksperimen 3: Perbandingan Arsitektur *Single-Instance* dan *Two-Stage* 61
4.8 Lingkungan Implementasi 61

V IMPLEMENTASI 62
5.1 Lingkungan Implementasi 62
5.2 *Libraries and Tools* 62
5.3 Implementasi Pra-pemrosesan Teks 65
5.4 Implementasi Ekstraksi dan Fusi Fitur 66
5.4.1 Representasi Leksikal 67
5.4.2 Representasi Semantik 67
5.4.3 Fitur Metadata 68
5.5 Implementasi Arsitektur *Single-Instance* 69
5.5.1 Rekonstruksi Laporan Tunggal 69
5.5.2 Pembentukan Vektor Fitur 70
5.5.3 Perbandingan dan Pemilihan *Classifier* 70
5.5.4 Penyetelan *Hyperparameter* 71
5.5.5 Pelatihan Akhir dan Evaluasi Kualitas 71
5.5.6 Pengukuran Efisiensi 72
5.6 Implementasi *Two-Stage* — *Stage* 1 (*Retrieval*) 72
5.6.1 Konstruksi Korpus dan Kueri 73
5.6.2 *Fine-tuning* SBERT 73
5.6.3 Pembentukan Sinyal Retrieval 74
5.6.4 Penggabungan Sinyal dan *Component Boost* 74
5.6.5 Ablasi Konfigurasi dan Pemilihan Final 75
5.6.6 Pembentukan Pasangan untuk *Stage* 2 76

UNIVERSITAS
GADJAH MADA
Deteksi Duplikasi Laporan Bug Berbasis Klasifikasi Menggunakan Arsitektur Single-Instance dan Two-Stage Pairwise
Muhammad Hariish Hafiiz, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

5.7 Implementasi *Two-Stage — Stage* 2 (*Pairwise Classifier*) 76
5.7.1 Ekstraksi Enam Fitur Relasional 77
5.7.2 Ablasi Modalitas 77
5.7.3 Perbandingan dan Pemilihan *Classifier* 78
5.7.4 Penyetelan *Hyperparameter* dan Pelatihan Akhir 78
5.8 Implementasi Pipeline *Two-Stage* Terpadu 79
5.8.1 Pemuatan Komponen Beku dan Klasifikasi Kandidat 79
5.8.2 Pemetaan ke Pasangan EP 80
5.8.3 Pengukuran Efisiensi 80
5.9 Implementasi *Baseline Pairwise* (Replikasi Ben Messaoud) 81
5.9.1 Representasi Pasangan dan *Tokenisasi* 81
5.9.2 Pelatihan Model 82
5.9.3 Pengukuran Biaya Inferensi $O(N)$ 83
5.10 Manajemen Artefak dan Reproduktibilitas 84
5.10.1 Penyimpanan Artefak pada HuggingFace 84
5.10.2 Konsistensi Model Antar Eksperimen 85
5.10.3 *Reproducibility* dan Pencegahan Kebocoran Data 85

VI HASIL DAN PEMBAHASAN 86
6.1 Analisis Dataset dan Karakteristik Data 86
6.1.1 Statistik Umum Dataset 86
6.1.2 Distribusi Pembagian Data 87
6.1.3 Distribusi Panjang Teks 87
6.1.4 Distribusi Metadata 88
6.1.5 Kontras Karakteristik Pasangan Duplikat dan Non-Duplikat 89
6.2 Hasil Arsitektur *Single-Instance* 91
6.2.1 Ablasi Modalitas 91
6.2.2 Perbandingan dan Pemilihan *Classifier* 93
6.2.3 Kinerja Akhir *Single-Instance* 94
6.2.4 Analisis *Feature Importance* 95
6.2.5 Ketahanan tanpa Fitur *Version* 96
6.2.6 Kinerja Komputasi 97
6.3 Hasil *Stage* 1 (*Retrieval*) 97
6.3.1 Hasil *Fine-tuning* SBERT 98
6.3.2 Perbandingan dan Ablasi Sinyal *Retrieval* 98

UNIVERSITAS
GADJAH MADA
Deteksi Duplikasi Laporan Bug Berbasis Klasifikasi Menggunakan Arsitektur Single-Instance dan Two-Stage Pairwise
Muhammad Hariish Hafiiz, Dr.techn. Guntur Budi Herwanto, S.Kom., M.Cs.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

6.3.3 Waktu Inferensi per Sinyal dan Analisis *Trade-off* 99
6.3.4 Pengaruh *Component Boost* 100
6.3.5 Kurva *Recall@K* dan Penetapan K 101
6.3.6 Analisis Penyebab Kegagalan *Retrieval* 102
6.3.7 Efisiensi *Retrieval* 103
6.4 Hasil *Stage* 2 (*Pairwise Classifier*) 104
6.4.1 Ablasi Modalitas 104
6.4.2 Perbandingan dan Pemilihan *Classifier* 105
6.4.3 Kinerja *Stage* 2 Terisolasi dan Perbandingan terhadap *Baseline* 107
6.4.4 Kontribusi Fitur *Version* pada *Stage* 2 108
6.4.5 Efisiensi *Stage* 2 109
6.5 Hasil *Pipeline Two-Stage End-to-End* 110
6.5.1 Cakupan *Retrieval* 110
6.5.2 Kinerja *End-to-End* 110
6.5.3 Pengaruh *Recall@K* terhadap Kualitas Akhir 111
6.6 Analisis Komparatif Tiga Arsitektur 113
6.6.1 Perbandingan Kualitas Deteksi 113
6.6.2 Perbandingan Efisiensi Komputasi 114
6.6.3 Analisis Pertukaran Kualitas dan Efisiensi 115
6.6.4 Posisi terhadap Literatur 116
6.7 Sintesis Temuan 117

VII KESIMPULAN DAN SARAN 120
7.1 Kesimpulan 120
7.2 Saran 121

DAFTAR PUSTAKA 122

LAMPIRAN 127