UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR ISI

HALAMAN PERSETUJUAN LAPORAN TESIS ii
HALAMAN PENGESAHAN iii
PRAKATA iv
DAFTAR ISI vi
DAFTAR GAMBAR ix
DAFTAR TABEL xii
INTISARI xv
ABSTRACT xvi
BAB I PENDAHULUAN 1
1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 3
1.3 Batasan Penelitian 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.6 Kontribusi Penelitian 5
BAB II TINJAUAN PUSTAKA 6
2.1 Klasifikasi Dialek Bahasa Arab 6
2.2 Ketidakseimbangan Kelas dalam Klasifikasi Dialek 7
2.3 Teknik *Oversampling* untuk Mengatasi Ketidakseimbangan Kelas 7
2.4 Sintesis dan Kebaruan Penelitian 9
BAB III METODOLOGI PENELITIAN 14
3.1 Dasar Operasional Metode 14
3.2 Kerangka Kerja Penelitian (*Research Framework*) 17
3.3 Data Penelitian: Shami Dialect Corpus 18
3.4 Prapemrosesan Teks 21
3.5 Rekayasa Fitur (*Feature Engineering*) dan Ekstraksi 24
3.6 Optimasi Model Dasar dan Pencarian Target *Oversampling* 28

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

3.7 Arsitektur Algoritma *Oversampling* yang Diusulkan 32
3.8 Skenario Pembobotan Sensitif Biaya (*Cost-Sensitive Learning*) 37
3.9 Pelatihan Model Klasifikasi 37
3.10 Rancangan Pengujian dan Evaluasi Model 40

BAB IV IMPLEMENTASI 46

4.1 Implementasi Prapemrosesan Teks Arab 46
4.2 Pemisahan Data dan Distribusi Kelas Asimetris 50
4.3 Rekayasa Fitur dan Pembangkitan Matriks Multiblok 53
4.4 Optimasi Model Dasar dan Target Penyeimbangan 60
4.5 Implementasi dan Optimasi Algoritma Usulan 63
4.6 Skenario Pembobotan Sensitif Biaya (*Cost-Sensitive Learning*) 65
4.7 Pelatihan Final Model Klasifikasi 67
4.8 Modul Pengujian dan Evaluasi 68

BAB V HASIL PENELITIAN DAN PEMBAHASAN 74

5.1 Verifikasi Hasil Pemrosesan Teks Arab 74
5.2 Hasil Pemisahan Data dan Distribusi Kelas Asimetris 74
5.3 Hasil Rekayasa Fitur dan Pembangkitan Matriks Multiblok 77
5.4 Hasil Optimasi Model Dasar dan Target Penyeimbangan 83
5.5 Hasil Implementasi dan Optimasi Algoritma Usulan 92
5.6 Kalkulasi Skenario Pembobotan Sensitif Biaya (*Cost-Sensitive Learning*) 113
5.7 Hasil Pelatihan Final Model 115
5.8 Hasil Pengujian dan Evaluasi Akhir 128
5.9 Perbandingan dengan penelitian terkini pada Shami Dialect Corpus 158

BAB VI KESIMPULAN DAN SARAN 164

6.1 Kesimpulan 164
6.2 Saran 165

DAFTAR PUSTAKA 168

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

LAMPIRAN 1. REPOSITORI ARTEFAK PENELITIAN, MATERI SUPLEMENTER, DEMO APLIKASI, DAN PUBLIKASI ILMIAH 177
LAMPIRAN 2. SURAT PERNYATAAN PLAGIARISME..... 178
LAMPIRAN 3. SURAT PERNYATAAN PENGGUNAAN TOOLS AI..... 179

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR GAMBAR

Gambar 3.1 Deskripsi umum sistem penelitian 18
Gambar 3.2 Rincian tahapan prapemrosesan data pada sistem klasifikasi 21
Gambar 3.3 Alur komputasi pencarian strategi penambahan data terbaik dan penguncian volume target berbasis F1-joint 31
Gambar 3.4 Alur kerja komputasional algoritma preventif ASTRA-SMOTE 35
Gambar 3.5 Alur kerja komputasional algoritma korektif SMOTE-RADIANT .. 36
Gambar 3.6 Proses pelatihan model klasifikasi 38
Gambar 3.7 Tahapan rancangan pengujian 42
Gambar 4.1 Kode program pembersihan derau non-linguistik 46
Gambar 4.2 Kode program pemangkasan elongasi *char stream* 48
Gambar 4.3 Kode program normalisasi ortografis bahasa Arab 48
Gambar 4.4 Kode program dua fungsi *analyzer* tokenisasi untuk *word stream* ... 49
Gambar 4.5 Kode program inisialisasi himpunan penyaringan entitas 50
Gambar 4.6 Kode program visualisasi distribusi kelas 51
Gambar 4.7 Kode *encoding* dan partisi data 52
Gambar 4.8 Kode program deklarasi aliran dan transformasi ruang leksikal 53
Gambar 4.9 Kode program ekstraksi fitur struktural dan morfologis 54
Gambar 4.10 Kode program *out-of-fold stacking* dan fitur interaksi probabilistik55
Gambar 4.11 Kode program penapisan dimensi fitur 57
Gambar 4.12 Kode program fusi horizontal blok fitur 58
Gambar 4.13 Kode program eksekusi transformasi fitur dan manajemen tembolok 59
Gambar 4.14 Kode program ruang pencarian parameter *fairness baseline* 60
Gambar 4.15 Kode program eksekusi pelatihan dan *early stopping* model dasar 61
Gambar 4.16 Kode program evaluasi metrik F1-joint 62
Gambar 4.17 Kode program iterasi pencarian volume *oversampling* 63
Gambar 4.18 Kode program optimasi parameter ASTRA-SMOTE 64
Gambar 4.19 Kode program optimasi parameter SMOTE-RADIANT 65
Gambar 4.20 Kode program inisialisasi bobot kelas 66

Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Gambar 4.21 Kode program eksekusi *oversampling* dengan konfigurasi optimal 67
Gambar 4.22 Kode program pelatihan final LightGBM dan XGBoost 67
Gambar 4.23 Kode program ekstraksi evaluasi prediksi dan analisis batas keputusan 69
Gambar 4.24 Kode program deteksi tingkat intrusi kelas sintetik 70
Gambar 4.25 Kode program agregasi kepentingan blok fitur 71
Gambar 4.26 Kode program iterasi uji *bootstrapping* dan Wilcoxon signed-rank test 72
Gambar 5.1 Visualisasi proporsi distribusi kelas asimetris pada Shami Dialect Corpus 75
Gambar 5.2 Perbandingan kepadatan jarak intrakelas sebelum dan sesudah SMOTE 90
Gambar 5.3 Proyeksi 2D t-SNE dan KDE *contour* (`perplexity=30`) sebelum dan sesudah SMOTE..... 91
Gambar 5.4 Grafik eksplorasi kombinatorial *hyperparameter* ASTRA-SMOTE pada mode Adaptive dan Full 93
Gambar 5.5 Komparasi kepadatan rata-rata intra-kelas antara Original, SMOTE, dan ASTRA-SMOTE 99
Gambar 5.6 Proyeksi 2D t-SNE dan KDE *contour* (`perplexity=30`) sebelum dan sesudah ASTRA-SMOTE 100
Gambar 5.7 Komparasi kepadatan rata-rata intra-kelas antara Original, SMOTE, dan SMOTE-RADIANT 107
Gambar 5.8 Proyeksi 2D t-SNE dan KDE *contour* (`perplexity=30`) sebelum dan sesudah SMOTE-RADIANT 108
Gambar 5.9 Proyeksi 2D t-SNE dan KDE *contour* (`perplexity=30`) sebelum dan sesudah ASTRA-RADIANT..... 112
Gambar 5.10 Kurva pembelajaran LightGBM pada Skenario 1 (Original)..... 116
Gambar 5.11 Kurva pembelajaran LightGBM pada Skenario 2 (SMOTE)..... 117
Gambar 5.12 Kurva pembelajaran LightGBM pada Skenario 3 (ASTRA-SMOTE) 118

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Gambar 5.13 Kurva pembelajaran LightGBM pada Skenario 4 (SMOTE-RADIANT) 119
Gambar 5.14 Kurva pembelajaran LightGBM pada Skenario 5 (ASTRA-RADIANT) 120
Gambar 5.15 Kurva pembelajaran LightGBM pada Skenario 6 (*class weight*) . 121
Gambar 5.16 Kurva pembelajaran XGBoost pada Skenario 1 (Original) 122
Gambar 5.17 Kurva pembelajaran XGBoost pada Skenario 2 (SMOTE) 123
Gambar 5.18 Kurva pembelajaran XGBoost pada Skenario 3 (ASTRA-SMOTE) 123
Gambar 5.19 Kurva pembelajaran XGBoost pada Skenario 4 (SMOTE-RADIANT) 124
Gambar 5.20 Kurva pembelajaran XGBoost pada Skenario 5 (ASTRA-RADIANT) 125
Gambar 5.21 Kurva pembelajaran XGBoost pada Skenario 6 (*class weight*) 126
Gambar 5.22 Perbandingan *confusion matrix* pada tiga skenario representatif.. 136
Gambar 5.23 Distribusi *confidence* prediksi benar dan salah pada tiga skenario representatif..... 138
Gambar 5.24 Distribusi *decision margin* menurut dialek dan ketepatan prediksi 139
Gambar 5.25 Distribusi entropi prediksi menurut dialek pada tiga skenario representatif..... 141
Gambar 5.26 Distribusi kepentingan blok fitur (*total importance*) LightGBM pada seluruh skenario 144
Gambar 5.27 Distribusi kepentingan blok fitur (*total importance*) XGBoost pada seluruh skenario 148
Gambar 5.28 Komparasi kinerja metrik klasifikasi antara model SOTA dan pendekatan usulan 161