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/

INTISARI

MODEL OVERSAMPLING HIBRIDA BERBASIS KLASTER DAN DENSITAS UNTUK PENANGANAN DATA TIDAK SEIMBANG PADA KASUS KLASIFIKASI DIALEK ARAB

Oleh

Maulana Ihsan Ahmad

24/551943/PPA/06991

Klasifikasi dialek Arab menghadapi tantangan berupa variasi linguistik dan ketidakseimbangan kelas ekstrem pada korpus media sosial. Penerapan SMOTE pada matriks teks berdimensi tinggi dan renggang dapat menghasilkan *overlapping noise* yang mengganggu pembentukan *decision boundary*. Penelitian ini mengembangkan dua model *hybrid oversampling*, yaitu ASTRA-SMOTE sebagai pendekatan preventif berbasis kemurnian klaster melalui MiniBatchKMeans dan SMOTE-RADIANT sebagai pendekatan korektif berbasis DBSCAN yang diterapkan pada sampel sintetik, serta menguji integrasi sekuensial keduanya dalam skenario ASTRA-RADIANT. Eksperimen dilakukan pada *Shami Dialect Corpus* menggunakan ruang fitur gabungan berdimensi 5.644, protokol bebas kebocoran secara ketat, serta arsitektur LightGBM dan XGBoost. Evaluasi mencakup performa klasifikasi, kualitas spasial sampel sintetik, analisis batas keputusan, dan pengujian *bootstrapping*. Hasil penelitian menunjukkan bahwa XGBoost dengan ASTRA-RADIANT menghasilkan *macro-F1* tertinggi sebesar 0,8546, sedangkan LightGBM dengan SMOTE-RADIANT menghasilkan *macro-recall* tertinggi sebesar 0,8469 dan *macro-F1* sebesar 0,8539. Dibandingkan SMOTE, ASTRA-RADIANT menurunkan *inter-class intrusion* dari 16,54% menjadi 1,05% pada Yordania, dari 22,04% menjadi 5,73% pada Lebanon, dan dari 9,83% menjadi 0,94% pada Palestina. Pengujian *bootstrapping* menunjukkan bahwa SMOTE-RADIANT dan ASTRA-RADIANT berbeda signifikan terhadap skenario Original pada kedua klasifikator, meskipun besar efeknya bergantung pada arsitektur. Temuan ini menunjukkan bahwa pengendalian pembangkitan dan penyaringan anomali dapat memperbaiki kualitas sampel sintetik serta keseimbangan pengenalan kelas minoritas, tetapi pemilihan metode tetap perlu disesuaikan dengan model dan prioritas metrik evaluasi.

Kata Kunci: klasifikasi dialek Arab, ketidakseimbangan kelas, ASTRA-SMOTE, SMOTE-RADIANT, ASTRA-RADIANT.

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/

ABSTRACT

CLUSTER AND DENSITY-BASED HYBRID OVERSAMPLING MODEL FOR HANDLING IMBALANCED DATA IN ARABIC DIALECT CLASSIFICATION

By

Maulana Ihsan Ahmad

24/551943/PPA/06991

Arabic dialect classification is challenged by linguistic variation and extreme class imbalance in social media corpora. Applying SMOTE to high-dimensional sparse text matrices can generate *overlapping noise* that disrupts decision-boundary formation. This study develops two hybrid oversampling models: ASTRA-SMOTE as a preventive cluster-purity approach using MiniBatchKMeans and SMOTE-RADIANT as a corrective DBSCAN-based approach applied to synthetic samples, and evaluates their sequential integration as ASTRA-RADIANT. Experiments were conducted on the *Shami Dialect Corpus* using a 5,644-dimensional fused feature space, a strict leakage-free protocol, and LightGBM and XGBoost architectures. The evaluation covered classification performance, the spatial quality of synthetic samples, decision-boundary analysis, and bootstrap testing. The results show that XGBoost with ASTRA-RADIANT achieved the highest *macro-F1* of 0.8546, while LightGBM with SMOTE-RADIANT achieved the highest *macro-recall* of 0.8469 and a *macro-F1* of 0.8539. Compared with SMOTE, ASTRA-RADIANT reduced inter-class intrusion from 16.54% to 1.05% for Jordanian, from 22.04% to 5.73% for Lebanese, and from 9.83% to 0.94% for Palestinian. Bootstrap testing showed that SMOTE-RADIANT and ASTRA-RADIANT differed significantly from the Original scenario for both classifiers, although the effect magnitude depended on the architecture. These findings indicate that controlling sample generation and filtering anomalies can improve synthetic-sample quality and minority-class recognition balance; however, method selection should remain conditional on the model and the prioritized evaluation metric.

Keywords: Arabic dialect classification, class imbalance, ASTRA-SMOTE, SMOTE-RADIANT, ASTRA-RADIANT.