

PERFORMA INDOBERT DENGAN OPTUNA PADA ANALISIS SENTIMEN DATA TIDAK SEIMBANG DI APLIKASI SIGNAL

ASFIAH ADIBA



**PROGRAM STUDI STATISTIKA DAN SAINS DATA
SEKOLAH SAINS, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Performa IndoBERT dengan Optuna pada Analisis Sentimen Data tidak Seimbang di Aplikasi Signal” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2025

Asfiah Adiba
G1401211004

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



ABSTRAK

ASFIAH ADIBA. Performa IndoBERT dengan Optuna pada Analisis Sentimen Data tidak Seimbang di Aplikasi Signal. Dibimbing oleh AKBAR RIZKI dan ANWAR FITRIANTO.

Analisis sentimen memainkan peran penting dalam memahami umpan balik pengguna terhadap aplikasi, khususnya dalam meningkatkan kualitas layanan digital. IndoBERT merupakan salah satu model pemrosesan bahasa alami terkemuka yang mampu memahami konteks teks secara mendalam. Penelitian ini menerapkan model IndoBERT pada aplikasi Signal – Samsat Digital Nasional, sebuah aplikasi pelayanan publik untuk pembayaran pajak kendaraan bermotor secara daring. Performa model dioptimalkan menggunakan *hyperparameter tuning* Optuna, yang terbukti lebih efisien dibandingkan *Random Search* maupun *Grid Search*. Selain itu, pendekatan penanganan data tidak seimbang juga diterapkan menggunakan teknik *oversampling* seperti SMOTE, *K-Means* SMOTE, dan ADASYN. Metrik evaluasi yang digunakan dalam penelitian ini adalah *balanced accuracy* dan *macro f1-score*, karena keduanya mampu memberikan penilaian performa yang lebih adil pada klasifikasi data tidak seimbang. Hasil terbaik diperoleh melalui kombinasi IndoBERT dengan *K-Means* SMOTE, yang menghasilkan nilai *balanced accuracy* sebesar 0,7623 dan *macro f1-score* sebesar 0,7846. Kombinasi ini meningkatkan *balanced accuracy* hingga 6% dan *macro f1-score* hingga 4% dibanding model tanpa penanganan data tidak seimbang. Hasil ini menunjukkan bahwa *K-Means* SMOTE merupakan teknik yang paling efektif di antara metode yang diuji dalam meningkatkan performa model pada tugas klasifikasi sentimen multi kelas dengan data yang tidak seimbang.

Kata kunci: ADASYN, IndoBERT, *K-Means* SMOTE, Signal, SMOTE



ABSTRACT

ASFIAH ADIBA. IndoBERT Performance with Optuna in Sentiment Analysis of Imbalanced Data on the Signal Application. Supervised by AKBAR RIZKI and ANWAR FITRIANTO.

Sentiment analysis is essential for understanding user feedback on applications. IndoBERT stands out as a leading model due to its ability to deeply understand text context. This study analyzes the latest version of the Signal – Samsat Digital Nasional application, a public service app for online motor vehicle tax payments. The IndoBERT model was optimized using Optuna hyperparameter tuning, which proved to be more efficient compared to Random Search or Grid Search. Additionally, approaches to handling imbalanced data were employed using *oversampling* technique such as SMOTE, K-Means SMOTE, and ADASYN. In this study, the evaluation metrics used are balanced accuracy and macro f1-score, as both provide a fairer assessment of model performance in imbalanced classification tasks. The best sentiment analysis results were achieved using IndoBERT with K-Means SMOTE for imbalanced data handling, yielding a balanced accuracy of 0,7623, macro f1-score of 0,7846. This combination improves balanced accuracy by up to 6% and macro f1-score by up to 4% compared to the baseline model without imbalance handling. These results highlight that K-Means SMOTE is the most effective technique among the tested methods for improving model performance in multiclass sentiment classification tasks with imbalanced data.

Keywords: ADASYN, IndoBERT, K-Means SMOTE, Signal, SMOTE



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

PERFORMA INDOBERT DENGAN OPTUNA PADA ANALISIS SENTIMEN DATA TIDAK SEIMBANG DI APLIKASI SIGNAL

ASFIAH ADIBA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi:
Prof. Dr. Ir. Hari Wijayanto, M.Si.



Judul Skripsi : Performa IndoBERT dengan Optuna pada Analisis Sentimen Data tidak Seimbang di Aplikasi Signal

Nama : Asfiah Adiba

NIM : G1401211004

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Akbar Rizki, S.Stat., M.Si.

Pembimbing 2:

Dr. Anwar Fitrianto, S.Si., M.Sc.

Diketahui oleh

Ketua Program Studi Statistika dan Sains Data:

Dr. Bagus Sartono, S.Si., M.Si.

NIP 197804112005011002

Tanggal Ujian:

9 Juli 2025

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Januari 2025 sampai bulan Juni 2025 ini ialah analisis sentimen pada aplikasi Signal, dengan judul “Performa IndoBERT dengan Optuna pada Analisis Sentimen Data tidak Seimbang di Aplikasi Signal”. Penulis ingin mengucapkan terima kasih kepada seluruh pihak yang telah berperan sejak penulis menempuh studi sampai proses penulisan karya ilmiah ini berakhir, di antaranya:

1. Ibu Akbar Rizki, S.Stat., M.Si. dan Bapak Dr. Anwar Fitrianto, S.Si., M.Sc. selaku komisi pembimbing skripsi yang telah memberikan arahan serta saran dalam penulisan karya ilmiah ini.
2. Ibu Aulia Rizki Firdawanti, S.Stat., M.Si. dan Bapak Ir. Aam Alamudi, M.Si. selaku moderator kolokium dan seminar yang telah memberikan masukan untuk perbaikan karya ilmiah ini.
3. Bapak Prof. Dr. Ir. Hari Wijayanto, M.Si. selaku dosen penguji yang telah memberikan masukan untuk perbaikan karya ilmiah ini.
4. Seluruh dosen dan staff akademik program studi Statistika dan Sains Data yang telah menyampaikan ilmunya dan membantu selama proses perkuliahan hingga menyelesaikan studi.
5. Seluruh keluarga besar penulis yang telah memberikan doa, dukungan, dan motivasi.
6. Seluruh teman khususnya Karimatu, Alfiah, Arfiah, Yasmin, Dhiya Kamilah, Kheni, Diva, Angel, Akmal, Adeva dan seluruh teman Statistika 58 yang telah menemani dan mewarnai masa perkuliahan.

Penulis menyadari bahwa karya ilmiah ini masih belum sempurna. Oleh karena itu, penulis memohon maaf apabila ditemukan kesalahan dan kekurangan dalam karya ilmiah ini. Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2025

Asfiah Adiba



DAFTAR ISI

DAFTAR TABEL	v
DAFTAR GAMBAR	v
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	2
II TINJAUAN PUSTAKA	3
2.1 Analisis Sentimen	3
2.2 Penanganan Data Tidak Seimbang	3
2.2.1 SMOTE	3
2.2.2 ADASYN	4
2.2.3 <i>K-Means</i> SMOTE	5
2.3 IndoBERT	6
2.4 <i>Hyperparameter Tuning</i> dengan Optuna	9
2.5 Evaluasi Performa IndoBERT	10
III METODE	12
3.1 Data	12
3.2 Prosedur Analisis Data	12
IV HASIL DAN PEMBAHASAN	15
4.1 Praproses Data	15
4.2 Pelabelan Data	18
4.3 Eksplorasi Data	20
4.4 Analisis Sentimen dengan IndoBERT	21
4.4.1 Persiapan Data Sebelum Pemodelan	21
4.4.2 Penanganan Data Tidak Seimbang	22
4.4.3 Pemodelan IndoBERT Tanpa Penanganan Data Tidak Seimbang	23
4.4.4 Pemodelan IndoBERT dengan SMOTE	26
4.4.5 Pemodelan IndoBERT dengan ADASYN	28
4.4.6 Pemodelan IndoBERT dengan <i>K-Means</i> SMOTE	31
4.4.7 Perbandingan Performa Model IndoBERT	33
V SIMPULAN DAN SARAN	36
5.1 Simpulan	36
5.2 Saran	36
DAFTAR PUSTAKA	37
RIWAYAT HIDUP	40

DAFTAR TABEL

1	Spesifikasi arsitektur dari BERT _{BASE} dan BERT _{LARGE}	8
2	Tampilan matiks konfusi multi kelas	11
3	Contoh data hasil <i>scraping</i>	12
4	Spesifikasi model IndoBERT	14
5	Contoh data sebelum dan setelah <i>case folding</i>	15
6	Contoh data sebelum dan setelah <i>text cleaning</i>	15
7	Contoh list KKBA untuk normalisasi kata	16
8	Contoh data sebelum dan setelah normalisasi kata	16
9	Distribusi data hasil pelabelan setiap metode pelabelan	19
10	Contoh data hasil pelabelan	19
11	Hasil setiap <i>trial</i> Optuna pada model IndoBERT tanpa penanganan data tidak seimbang	24
12	<i>Classification report</i> dari model IndoBERT tanpa penanganan data tidak seimbang	25
13	Hasil setiap <i>trial</i> Optuna pada model IndoBERT SMOTE	27
14	<i>Classification report</i> dari model IndoBERT SMOTE	27
15	Hasil setiap <i>trial</i> Optuna pada model IndoBERT ADASYN	29
16	<i>Classification report</i> dari model IndoBERT ADASYN	30
17	Hasil setiap <i>trial</i> Optuna pada model IndoBERT <i>K-Means</i> SMOTE	31
18	<i>Classification report</i> dari model IndoBERT <i>K-Means</i> SMOTE	32
19	Perbandingan performa model IndoBERT	34

DAFTAR GAMBAR

1	Ilustrasi SMOTE (Joloudari <i>et al.</i> 2023)	4
2	Ilustrasi ADASYN (Elnahas <i>et al.</i> 2022)	4
3	Ilustrasi <i>K-Means</i> SMOTE (Last <i>et al.</i> 2017)	6
4	Arsitektur model BERT (Devlin <i>et al.</i> 2019)	7
5	Ilustrasi representasi input BERT (Devlin <i>et al.</i> 2019)	8
6	<i>Flowchart</i> prosedur analisis	14
7	Distribusi jumlah data berdasarkan ulasan bintang sebelum dan setelah penghapusan data duplikat	17
8	<i>Pie chart</i> distribusi data setelah penghapusan data duplikat	17
9	<i>Pie chart</i> sebaran jumlah data tiap kelas	19
10	<i>Word clouds</i> pada setiap kelas (a) positif, (b) netral, (c) negatif	20
11	Distribusi (a) data latih, (b) data validasi, (c) data uji	21
12	<i>Bar chart</i> distribusi data sebelum dan setelah <i>oversampling</i>	22
13	Matiks konfusi model IndoBERT tanpa penanganan data tidak seimbang pada data uji	25
14	Matiks konfusi model IndoBERT SMOTE pada data uji	28
15	Matiks konfusi model IndoBERT ADASYN pada data uji	30
16	Matiks konfusi model IndoBERT <i>K-Means</i> SMOTE pada data uji	33
17	Plot akurasi (a) dan <i>loss</i> (b) selama pelatihan dan validasi	35