EVALUASI DAMPAK RANDOM OVER SAMPLING TERHADAP KINERJA INDOBERT UNTUK ANALISIS SENTIMEN BAHASA INDONESIA

SKRIPSI NON REGULER JALUR SCIENTIST

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

Disusun oleh :

DIMAS RAMADHAN ALFINSYAH

22.11.4742

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

EVALUASI DAMPAK RANDOM OVER SAMPLING TERHADAP KINERJA INDOBERT UNTUK ANALISIS SENTIMEN BAHASA INDONESIA

SKRIPSI NON REGULER JALUR SCIENTIST

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

Disusun oleh :

DIMAS RAMADHAN ALFINSYAH

22.11.4742

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

JALUR NON-REGULER

EVALUASI DAMPAK RANDOM OVER SAMPLING TERHADAP KINERJA INDOBERT UNTUK ANALISIS SENTIMEN BAHASA INDONESIA

yang disusun dan diajukan oleh

DIMAS RAMADHAN ALFINSYAH

22.11.4742

telah disetujui oleh Dosen Pembimbing

pada tanggal 17 Desember 2025

Dosen Pembimbing,

Bambang Pilu Hartato, S.Kom., M.Eng.
NIK. 190302707

HALAMAN PENGESAHAN

JALUR NON-REGULER

EVALUASI DAMPAK RANDOM OVER SAMPLING TERHADAP KINERJA INDOBERT UNTUK ANALISIS SENTIMEN BAHASA INDONESIA

yang disusun dan diajukan oleh

DIMAS RAMADHAN ALFINSYAH

22.11.4742

Telah dipertahankan di depan Dewan Penguji

pada tanggal 17 Desember 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Uyock Anggoro Saputro, S.Kom., M.Kom. **NIK. 190302419**	
Windha Mega Pradnya Dhuhita, S.Kom., M.Kom. **NIK. 190302185**	
Bambang Pilu Hartato, S.Kom., M.Eng. **NIK. 190302707**	

Laporan ini telah diterima sebagai salah satu persyaratan

untuk memperoleh gelar Sarjana Komputer

Tanggal 17 Desember 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.

NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN KARYA

Yang bertandatangan di bawah ini,

Nama mahasiswa : **DIMAS RAMADHAN ALFINSYAH**

NIM : **22.11.4742**

Menyatakan bahwa Laporan dengan judul berikut:

Evaluasi Dampak Random Over Sampling Terhadap Kinerja Indobert Untuk Analisis Sentimen Bahasa Indonesia

Dosen Pembimbing : **Bambang Pilu Hartato, S.Kom., M.Eng.**

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan kegiatan SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidak-benaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 17 Desember 2025

Yang Menyatakan,

Dimas Ramadhan Alfinsyah

HALAMAN PERSEMBAHAN

Puji syukur penulis panjatkan ke hadirat **Allah SWT** atas rahmat, taufik, dan hidayah-Nya sehingga laporan ini dapat terselesaikan dengan baik. Penyusunan laporan ini tidak lepas dari bantuan, dukungan, serta doa dari berbagai pihak yang senantiasa memberikan semangat dan motivasi. Dengan penuh rasa hormat, laporan ini penulis persembahkan kepada:

1. **Kedua orang tua tercinta**, Bapak Marjadi dan Ibu Ika Kumarasari , serta kakak Desta, yang selalu memberikan doa, kasih sayang, dan dukungan tanpa henti. Terima kasih atas segala pengorbanan dan kerja keras yang telah diberikan agar penulis dapat menempuh pendidikan hingga tahap ini.
2. **Bapak/Ibu dosen pembimbing dan penguji**, yang telah membimbing dan memberikan arahan dengan penuh kesabaran serta memberikan masukan berharga demi kesempurnaan laporan ini.
3. **Teman-teman seperjuangan** di kelas S1-Informatika 03, yang selalu memberikan semangat, kebersamaan, dan dukungan selama masa perkuliahan hingga penyusunan laporan ini. Terima kasih atas tawa, bantuan, dan kebersamaan yang telah menjadi bagian dari perjalanan ini.
4. **Semua pihak** yang telah membantu secara langsung maupun tidak langsung. Terima kasih atas doa dan dukungannya hingga laporan ini dapat terselesaikan dengan baik.

Akhir kata, semoga laporan ini dapat memberikan manfaat, menambah wawasan, serta menjadi kontribusi kecil bagi pengembangan ilmu pengetahuan di lingkungan akademik AMIKOM.

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat **Tuhan Yang Maha Esa** atas segala rahmat, karunia, dan petunjuk-Nya sehingga penulis dapat menyelesaikan **Laporan Skripsi Non Reguler Jalur Scientist** yang berjudul "Evaluasi Dampak Random Over Sampling Terhadap Kinerja Indobert Untuk Analisis Sentimen Bahasa Indonesia".

Penyusunan laporan ini merupakan salah satu syarat untuk menyelesaikan program studi **S1 Informatika** di **Universitas Amikom Yogyakarta**. Laporan ini disusun sebagai hasil dari penelitian dan pembelajaran yang telah penulis jalani selama masa perkuliahan, khususnya dalam bidang keilmuan yang berkaitan dengan topik penelitian ini.

Dalam proses penyusunan laporan ini, penulis banyak menerima bantuan, dukungan, dan bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. **Allah SWT**, atas limpahan rahmat dan karunia-Nya sehingga laporan ini dapat terselesaikan dengan baik.
2. **Bapak Prof. Dr. M. Suyanto, M.M.**, selaku Rektor Universitas Amikom Yogyakarta.
3. **Ibu Eli Pujastuti, M.Kom.**, selaku Ketua Program Studi S1 Informatika. Universitas Amikom Yogyakarta.
4. **Bapak Bambang Pilu Hartato, S.Kom., M.Eng.**, selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta motivasi selama proses penyusunan laporan ini.
5. **Seluruh dosen dan staf Universitas Amikom Yogyakarta**, yang telah memberikan ilmu dan bantuan administrasi selama penulis menempuh pendidikan.

6. **Kedua orang tua dan keluarga tercinta**, yang selalu memberikan doa, dukungan moral, dan semangat tanpa henti.
7. **Rekan-rekan dan sahabat seperjuangan**, yang senantiasa memberikan semangat, bantuan, serta kebersamaan selama proses penyusunan laporan ini.

Penulis menyadari bahwa laporan ini masih jauh dari sempurna karena keterbatasan pengetahuan dan pengalaman. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun demi perbaikan di masa mendatang.

Akhir kata, semoga laporan ini dapat memberikan manfaat bagi penulis sendiri maupun bagi pembaca yang tertarik dengan topik yang dibahas.

Yogyakarta, 6 Desember 2025

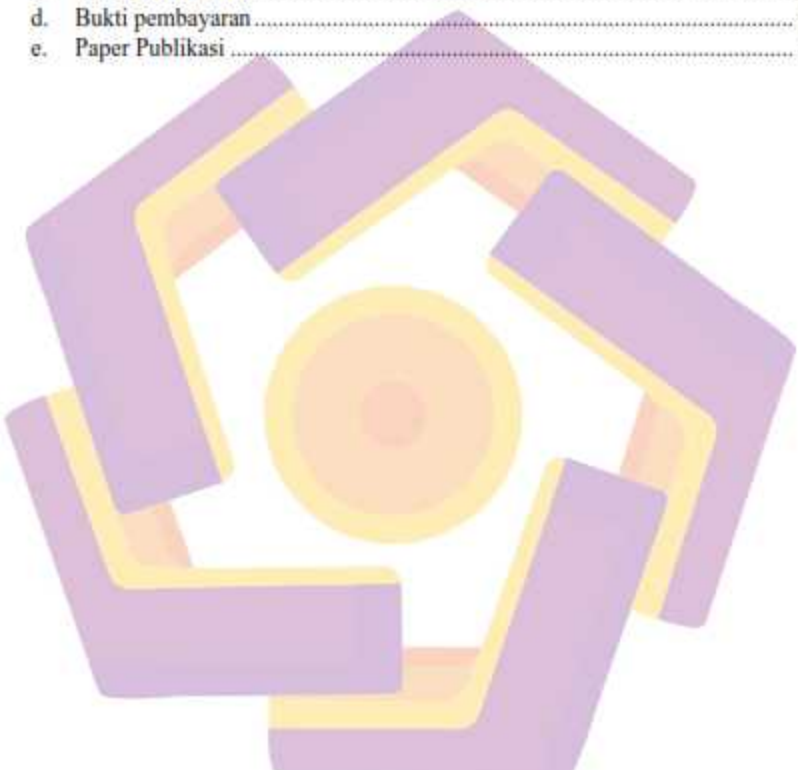
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

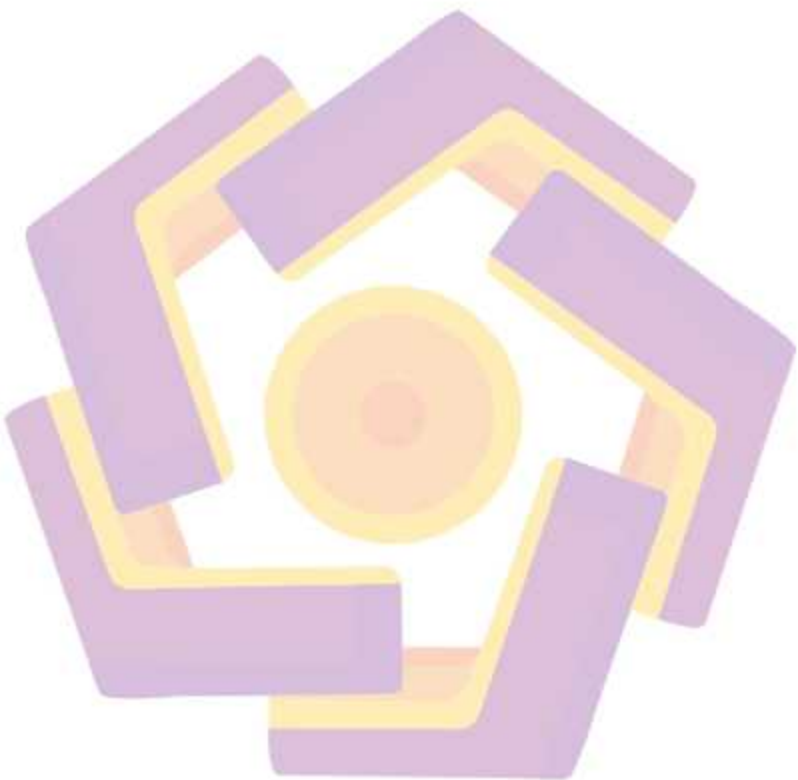
DAFTAR LAMPIRAN



DAFTAR LAMBANG DAN SINGKATAN

ACC	Accuracy
ADASYN	Adaptive Synthetic Sampling
AdamW	Adaptive Moment Estimation with Weight Decay
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BERTForSequenceClassification	Arsitektur model klasifikasi berbasis BERT
CLS	Classification token (token awal pada model BERT)
COLAB	Google Colaboratory
CSV	Comma Separated Values
F1	F1-Score
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
ID	Identifier
IDN App	Aplikasi berita digital "IDN" (konteks penelitian)
IndoBERT	Indonesian Bidirectional Encoder Representations from Transformers
LDA	Latent Dirichlet Allocation
MBTI	Myers–Briggs Type Indicator
NLP	Natural Language Processing
GPU NVIDIA T4	Jenis Graphics Processing Unit dari NVIDIA tipe T4
PREC	Precision
REC	Recall
ROS	Random Over Sampler / Random Over Sampling
RUS	Random Under Sampling
SEP	Separator token (penanda akhir urutan pada model BERT)
SMOTE	Synthetic Minority Over-sampling Technique
TN	True Negative

TP	True Positive



DAFTAR ISTILAH

Accuracy	Persentase prediksi benar
Analisis Sentimen	Proses menentukan sentimen positif atau negatif dalam teks
Attention Mechanism	Mekanisme fokus model pada token penting
Attention Mask	Penanda token mana yang diproses oleh model
Binary Classification	Klasifikasi dua kelas
Case Folding	Mengubah huruf menjadi bentuk kecil (seragam)
Class Weighting	Teknik pembobotan agar kelas minoritas berpengaruh lebih besar saat pelatihan
Colloquial Lexicon	Kamus untuk normalisasi kata tidak baku ke bentuk formal
Confusion Matrix	Tabel evaluasi hasil klasifikasi
Cost Sensitive Learning	Metode yang memberi penalti lebih besar pada kesalahan di kelas minoritas.
Cross Entropy Loss	Mengukur perbedaan antara distribusi prediksi model dan label sebenarnya
Data Augmentation	Penambahan variasi data
DataFrame	Struktur data tabel di Pandas
DataLoader	Pemuat data dalam batch pada PyTorch
Dataset	Kumpulan data untuk pelatihan model
Early Stopping	Menghentikan pelatihan sebelum overfitting
Epoch	Satu siklus penuh pelatihan model pada seluruh dataset
F1-Score	Rata-rata harmonik precision dan recall
Fine-tuning	Penyesuaian model pre-trained
Inference	Proses penggunaan model yang telah dilatih untuk memprediksi data baru
Ketidakseimbangan Data	Jumlah data antar kelas tidak seimbang

Labeling	Pemberian label kelas pada data
Learning Curve	Grafik yang menunjukkan perkembangan performa model selama pelatihan
Learning Rate	Kecepatan perubahan bobot model saat pelatihan
Library	Kumpulan fungsi atau modul siap pakai dalam pemrograman
Noise	Data acak atau tidak relevan yang dapat mengganggu proses pelatihan model
Normalisasi	Mengubah kata tidak baku menjadi baku
Overfitting	Model terlalu cocok pada data latih
Over sampling	Menambah sampel pada kelas minoritas untuk menyeimbangkan data.
Padding	Menambahkan token kosong agar panjang urutan seragam
Pipeline	Rangkaian proses atau alur kerja otomatis dalam sistem pemrosesan data
Precision	Ketepatan prediksi positif
Preprocessing	Tahap pembersihan dan persiapan data
Pseudocode	Representasi langkah algoritma dalam bentuk mirip kode
Punctuation Removal	Menghapus tanda baca yang tidak perlu
Recall	Kemampuan menangkap data positif
Regularisasi	Pencegahan overfitting pada model
Resampling	Menyeimbangkan jumlah data antar kelas
Sastrawi Stemmer	Algoritma stemming untuk Bahasa Indonesia
Scheduler	Pengatur perubahan learning rate selama training
Stemming	Mengubah kata berimbuhan ke bentuk dasar
Stopword	Kata umum yang diabaikan dalam analisis
Tensor	Representasi data multidimensi
Test Set	Data untuk pengujian akhir model

Tokenisasi	Memecah teks menjadi kata atau token
Training Set	Data yang digunakan untuk melatih model
Transformer	Arsitektur model berbasis attention
Truncating	Memotong urutan teks yang terlalu panjang
Under sampling	Mengurangi sampel pada kelas mayoritas agar data seimbang.
Optimizer	Algoritma untuk memperbarui bobot model selama training
Validation Set	Data untuk evaluasi selama pelatihan
Web Scraping	Teknik otomatis mengambil data dari halaman web
WordPiece Tokenizer	Algoritma pemecah kata menjadi subword

INTISARI

Analisis sentimen merupakan salah satu bidang penelitian penting dalam *Natural Language Processing (NLP)*. Untuk bahasa Indonesia, *IndoBERT* telah muncul sebagai model unggulan berkat kinerjanya yang kompetitif. Namun, efektivitasnya sangat dipengaruhi oleh distribusi kelas yang seimbang. Tantangan umum muncul karena ulasan pengguna pada platform digital, seperti *Google Play Store*, sering kali menunjukkan ketidakseimbangan kelas. Penelitian ini mengkaji efektivitas teknik *Random Over Sampler (ROS)* dalam meningkatkan kinerja *IndoBERT* pada kondisi data yang tidak seimbang. Dataset yang digunakan terdiri dari 13.821 ulasan pengguna aplikasi *IDN App* yang dikumpulkan dari *Google Play Store* sejak tahun 2015 hingga Juli 2025. Sebelum proses pemodelan, dilakukan tahap praproses data yang meliputi penghapusan tanda baca, *case folding*, penghapusan *stopword*, *tokenizing*, normalisasi, dan *stemming* untuk memastikan konsistensi teks. Ulasan kemudian dikategorikan ke dalam dua kelas sentimen, yaitu positif (rating 3–5 bintang) dan negatif (rating 1–2 bintang). Dua skenario eksperimen dilakukan, yaitu (1) *IndoBERT* tanpa *ROS* dan (2) *IndoBERT* dengan dataset seimbang menggunakan *ROS*. Kinerja model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, dengan pembagian data sebesar 70% untuk pelatihan, 20% untuk validasi, dan 10% untuk pengujian. Hasil penelitian menunjukkan adanya peningkatan signifikan setelah penerapan *ROS*, yaitu *accuracy* sebesar 94,55%, *precision* 94,61%, *recall* 94,53%, dan *F1-score* 94,54%. Analisis *confusion matrix* menunjukkan peningkatan dalam klasifikasi kelas minoritas dengan penurunan tingkat kesalahan hingga 49%. Namun, analisis *learning curve* mengindikasikan adanya potensi *overfitting* akibat penggunaan *ROS*. Oleh karena itu, penelitian lanjutan diperlukan untuk mengoptimalkan strategi *ROS* agar dapat mencapai kinerja dan kemampuan generalisasi yang lebih baik.

Kata kunci: Analisis Sentimen, IndoBERT, *Random Over Sampler*, Data Tidak Seimbang, Evaluasi Model.

ABSTRACT

Sentiment analysis is a prominent research area in natural language processing (NLP). For the Indonesian language, IndoBERT has emerged as a leading model due to its competitive performance. However, its effectiveness is strongly influenced by balanced class distribution. A common challenge arises because user reviews on digital platforms, such as the Google Play Store, often exhibit imbalanced classes. This study investigates the effectiveness of the Random Over Sampler (ROS) technique in improving IndoBERT's performance under imbalanced data conditions. The dataset consists of 13,821 user reviews of the IDN App collected from the Google Play Store between 2015 and July 2025. Prior to modeling, data preprocessing was performed, including punctuation removal, case folding, stopword removal, tokenizing, normalization, and stemming to ensure textual consistency. Reviews were categorized into two sentiment classes: positive (3–5 stars) and negative (1–2 stars). Two experimental scenarios were conducted: (1) IndoBERT without ROS and (2) IndoBERT with a balanced dataset using ROS. Model performance was evaluated using accuracy, precision, recall, and F1-score, with data split into 70% training, 20% validation, and 10% testing. Results showed a significant improvement after ROS implementation: 94.55% accuracy, 94.61% precision, 94.53% recall, and 94.54% F1-score. Confusion matrix analysis indicated improved classification of the minority class, reducing the error rate by 49%. However, learning curve analysis revealed potential overfitting due to ROS. Further research is needed to optimize ROS strategies for better performance and generalization.

Keywords: *Sentiment Analysis, IndoBERT, Random Over Sampler, Imbalanced Data, Model Evaluation.*