

ANALISIS KOMPARATIF MBERT DAN INDOBERT UNTUK KLASIFIKASI SENTIMEN TWEET BERBAHASA INDONESIA INFORMAL DAN CODE-MIXING

TESIS



**PROGRAM STUDI ILMU KOMPUTER (S2)
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA**



- UU ITE No. 11 Tahun 2008 Pasal 5 Ayat 1 "Informasi Elektronik dan/atau hasil cetaknya merupakan alat bukti hukum yang sah"
- Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan BSrE - BSSN, validitas dokumen elektronik ini bisa dicek menggunakan aplikasi mobile VeryDS oleh BSrE
- Cetak dokumen ini merupakan salinan dari file dokumen bertandatangan elektronik yang keabsahannya dapat diakses melalui scan QRCode yang terdapat pada sertifikat ini.



- UU ITE No. 11 Tahun 2008 Pasal 5 Ayat 1 "Informasi Elektronik dan/atau hasil cetaknya merupakan alat bukti hukum yang sah"
- Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan BSrE - BSSN, validitas dokumen elektronik ini bisa dicek menggunakan aplikasi mobile VeryDS oleh BSrE
- Cetak dokumen ini merupakan salinan dari file dokumen bertandatangan elektronik yang keabsahannya dapat diakses melalui scan QRCode yang terdapat pada sertifikat ini.



Balai
Sertifikasi
Elektronik

TESIS

DIAJUKAN UNTUK MELENGKAPI TUGAS DAN MEMENUHI SYARAT-SYARAT UNTUK MENCAPAI GELAR MAGISTER KOMPUTER

Menyetujui

Pembimbing I	Dr. I Gede Aris Gunadi, S.Si. M.Kom. NIP.197703182008121004
Pembimbing II	Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs. NIP.198307252008011008



- UU ITE No. 11 Tahun 2008 Pasal 5 Ayat 1 "Informasi Elektronik dan/atau hasil cetaknya merupakan alat bukti hukum yang sah"
- Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan BSrE - BSSN, validitas dokumen elektronik ini bisa dicek menggunakan aplikasi mobile VeryDS oleh BSrE
- Cetak dokumen ini merupakan salinan dari file dokumen bertandatangan elektronik yang keabsahannya dapat diakses melalui scan QRCode yang terdapat pada sertifikat ini.

Tesis oleh I Komang Adyanata telah dipertahankan di depan tim penguji dan dinyatakan diterima sebagai salah satu persyaratan untuk memperoleh gelar Magister Komputer di Ilmu Komputer (S2), Program Pascasarjana, Universitas Pendidikan Ganesha
Pada tanggal 03 Agustus 2026

Dewan Penguji

Ketua	Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs. NIP.198307252008011008
Anggota	Dr. I Gede Aris Gunadi, S.Si. M.Kom. NIP.197703182008121004
Anggota	Prof. Dr. Gede Suweken, M.Sc. NIP.196111111987021001
Anggota	Prof. Drs. Sariyasa, M.Sc., Ph.D. NIP.196406151989021001

Mengetahui Direktur Program Pascasarjana
Universitas Pendidikan Ganesha,



Prof. Dr. I Nyoman Jampel, M.Pd.

NIP.195910101986031003



- UU ITE No. 11 Tahun 2008 Pasal 5 Ayat 1 "Informasi Elektronik dan/atau hasil cetaknya merupakan alat bukti hukum yang sah"
- Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan BSrE - BSSN, validitas dokumen elektronik ini bisa dicek menggunakan aplikasi mobile VeryDS oleh BSrE
- Cetakan dokumen ini merupakan salinan dari file dokumen bertandatangan elektronik yang keabsahannya dapat diakses melalui scan QRCode yang terdapat pada sertifikat ini.

LEMBAR PERNYATAAN

Saya menyatakan dengan sesungguhnya bahwa tesis yang saya susun sebagai syarat untuk memperoleh gelar Magister Ilmu Komputer dari Pascasarjana Universitas Pendidikan Ganesha seluruhnya merupakan hasil karya saya sendiri. Bagian-bagian tertentu dalam penulisan tesis yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya secara jelas dan sesuai dengan norma, kaidah, serta etika akademis.

Apabila di kemudian hari ditemukan seluruh atau sebagian tesis ini bukan hasil karya saya sendiri atau adanya plagiat dalam bagian-bagian tertentu, saya bersedia menerima sanksi pencabutan gelar akademik yang saya sandang dan sanksi-sanksi lainnya sesuai dengan peraturan perundang-undangan yang berlaku di wilayah Negara Kesatuan Republik Indonesia.

Singaraja, 3 Agustus 2026

Yang memberi pernyataan,



I Komang Adyanata

PRAKATA

Om Swastyastu

Puji dan syukur dipanjatkan ke hadapan Ida Sang Hyang Widhi Wasa, Tuhan Yang Maha Esa, atas asung kerta wara nugraha-Nya, Tesis yang berjudul “Analisis Komparatif mBERT Dan IndoBERT untuk Klasifikasi Sentimen *Tweet* Berbahasa Indonesia Informal dan *Code-Mixing*” ini dapat diselesaikan sesuai dengan yang direncanakan.

Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Ilmu Komputer pada Program Studi Ilmu Komputer, Pascasarjana, Universitas Pendidikan Ganesha. Proses penyusunan tesis ini tidak lepas dari berbagai tantangan yang pada akhirnya dapat penulis lalui berkat bantuan, bimbingan, dorongan, serta doa dari berbagai pihak. Dengan penuh rasa syukur dan kerendahan hati, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

- Dr. I Gede Aris Gunadi, S.Si. M.Kom., selaku Pembimbing I sekaligus sebagai Sekretaris Program Studi Ilmu Komputer Pascasarjana Universitas Pendidikan Ganesha, yang dengan penuh kesabaran dan dedikasi tinggi telah meluangkan waktu untuk memberikan bimbingan, arahan, dan motivasi yang sangat berarti dari awal hingga selesainya tesis ini;
- Dr. I Made Gede Sunarya, S.Kom., M.Cs., selaku Pembimbing II sekaligus sebagai Koordinator Program Studi Ilmu Komputer Pascasarjana Universitas Pendidikan Ganesha, atas masukan yang konstruktif dan sangat membantu dalam menyempurnakan isi dan sistematika penulisan tesis ini;
- Prof. Dr. Gede Suweken, M.Sc. dan Prof. Drs. Sariyasa, M.Sc., Ph.D., selaku Dewan Penguji yang telah memberikan kritik dan saran yang sangat membangun untuk tesis ini;
- Direktur Pascasarjana Universitas Pendidikan Ganesha dan staf, yang telah banyak membantu selama penulis menyelesaikan tesis ini;
- Rektor Universitas Pendidikan Ganesha, yang telah memberikan bantuan secara moral dan memfasilitasi berbagai kepentingan penulis dalam menyelesaikan tesis ini;

- Orang tua penulis (Bapak dan Ibu) serta Istri tercinta, yang telah memberikan segalanya, doa, kasih sayang, dukungan moral dan materiil kepada penulis;
- Sahabat dan rekan-rekan mahasiswa Program Studi Ilmu Komputer Pascasarjana Universitas Pendidikan Ganesha yang telah menjadi teman diskusi, berbagi semangat, dan saling menguatkan selama masa perkuliahan hingga penyelesaian tesis ini;
- Serta semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan bantuan dan dukungan baik secara langsung maupun tidak langsung dalam penyelesaian tesis ini.

Penulis menyadari sepenuhnya bahwa tesis ini masih jauh dari sempurna. Namun, kehadirannya dalam konstelasi masyarakat akademis akan menambah perbendaharaan ilmu dalam perkembangan ilmu pengetahuan. Akhir kata, semoga Ida Sang Hyang Widhi Wasa senantiasa melimpahkan asung kerta wara nugraha-Nya kepada semua pihak yang telah memberikan kontribusi dalam penyelesaian tesis ini. Semoga tesis ini dapat memberikan manfaat nyata bagi pengembangan ilmu pengetahuan di bidang Ilmu Komputer serta menjadi referensi yang berguna bagi peneliti selanjutnya.

Om Shanti, Shanti, Shanti Om

Singaraja, Agustus 2026

Penulis

DAFTAR ISI

(halaman)

LEMBAR PERNYATAAN	v
PRAKATA	vi
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xiii
DAFTAR LAMPIRAN	xiv
ABSTRAK	xv
ABSTRACT	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang Penelitian	1
1.2 Batasan Masalah	3
1.3 Rumusan Masalah	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	6
BAB II KAJIAN PUSTAKA	7
2.1 Kajian Teori	7
2.1.1 Media Sosial sebagai Sumber Data Sentimen	7
2.1.2 Analisis Sentimen dan <i>Opinion mining</i>	8
2.1.3 Arsitektur <i>Transformer</i>	9
2.1.4 <i>Bidirectional Encoder Representations from Transformers (BERT)</i> ...	10
2.1.5 <i>Fine-tuning BERT</i>	14
2.1.6 <i>Multilingual BERT (mBERT)</i>	15
2.1.7 <i>IndoBERT</i>	16
2.1.8 <i>Support Vector Machine (SVM)</i>	17
2.1.9 <i>Term frequency-Inverse Document frequency (TF-IDF)</i>	18
2.1.10 <i>Audit Tokenizer</i>	19
2.1.11 <i>K-fold Cross validation</i>	20
2.1.12 <i>Confusion Matrix</i> dan Evaluasi	20
2.1.13 <i>Hyperparameter Tuning</i>	22
2.1.14 <i>Class weighting</i>	25
2.1.15 Uji <i>McNemar</i> dan <i>Bootstrap Confidence interval</i>	26
2.2 Kajian Hasil Penelitian yang Relevan	27
2.2.1 Pengembangan Model <i>Pre-trained</i> untuk Bahasa Indonesia	27
2.2.2 Aplikasi <i>Transformer</i> pada Sentimen Twitter Indonesia	28

2.2.3 Perbandingan Metode Klasik dan <i>Transformer</i>	30
2.2.4 Penanganan Ketidakseimbangan Kelas pada Klasifikasi Teks	32
2.3 Hipotesis Penelitian.....	33
BAB III METODE PENELITIAN.....	35
3.1 Teknik Pengumpulan Data	35
3.1.1 Studi Literatur	35
3.1.2 Data <i>Tweet</i> Sosial Media Twitter	35
3.2 Tempat dan Waktu Penelitian	36
3.3 Tahapan Penelitian	36
3.3.1 <i>Preprocessing</i> Data	37
3.3.2 Pemformatan Data.....	38
3.3.3 Peninjauan Pelabelan	39
3.3.4 <i>Exploratory Data Analysis</i>	39
3.3.5 Implementasi mBERT dan IndoBERT	40
3.3.6 <i>Hyperparameter Tuning</i> Dua Tahap	40
3.3.7 Eksekusi <i>Baseline TF-IDF + Linear SVM</i>	42
3.3.8 Audit Komprehensif <i>Tokenizer</i>	42
3.3.9 Uji Coba dan Evaluasi	43
3.3.10 Dokumentasi	44
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	46
4.1 Karakteristik Indonesian General <i>Sentiment Analysis Dataset</i>	46
4.1.1 Distribusi Label Sentimen	47
4.1.2 Statistik Panjang <i>Tweet</i>	48
4.1.3 Analisis Penggunaan Bahasa Informal	49
4.2 Perbandingan Proses Tokenisasi mBERT dan IndoBERT	50
4.2.1 Tokenisasi pada Sampel Kata Tidak Baku.....	51
4.2.2 Statistik Fragmentasi <i>Subword</i>	52
4.2.3 Ilustrasi Tokenisasi pada <i>Tweet</i> Lengkap	53
4.2.4 Audit Komprehensif <i>Tokenizer</i>	53
4.2.5 Contoh Hasil <i>Token</i> dari <i>Dataset</i>	61
4.3 <i>Hyperparameter Tuning</i> dan Evaluasi Performa Klasifikasi	71
4.3.1 Hasil Pencarian <i>Hyperparameter</i> Tahap Pertama	71
4.3.2 Eksplorasi <i>Epoch</i> Penuh dan Konfigurasi Final.....	72
4.3.3 Hasil Evaluasi pada <i>Test set</i>	73
4.3.4 Perbandingan dengan <i>Baseline</i> Penelitian Sebelumnya.....	74
4.3.5 Analisis Per Kelas dan <i>Confusion matrix</i>	75
4.3.6 <i>Baseline TF-IDF + SVM</i> sebagai Pembanding Klasik.....	76
4.3.7 Ablasi <i>Class weighting</i>	81
4.3.8 Analisis Potensi <i>Overfitting</i>	82
4.4 Pengaruh Karakteristik <i>Dataset</i> dan Tokenisasi terhadap Performa.....	86
4.4.1 Performa per Kategori Karakteristik	86

4.4.2 Korelasi antara Fragmentasi Tokenisasi dan <i>Accuracy</i>	88
4.4.3 Studi Kasus <i>Disagreement</i> antar Model.....	89
4.4.4 Analisis Pola dan Penyebab Kesalahan Klasifikasi	89
4.5 Perbandingan IndoBERT dan mBERT Sesuai Hasil Penelitian	93
4.5.1 Uji Signifikansi Statistik atas Perbedaan Performa	95
BAB V SIMPULAN DAN SARAN	99
5.1 Simpulan	99
5.2 Saran.....	101
DAFTAR PUSTAKA	103
LAMPIRAN.....	110



DAFTAR TABEL

	<i>(halaman)</i>
Tabel 2.1 Contoh <i>Confusion Matrix</i> untuk <i>multi-class classification</i> .	21
Tabel 3.1 Kesetaraan Prosedur Tuning antara mBERT dan IndoBERT	41
Tabel 4.1 Distribusi Label Sentimen pada <i>Dataset</i>	47
Tabel 4.2 Statistik Deskriptif Panjang <i>Tweet</i> (Setelah Pra-pemrosesan)	48
Tabel 4.3 Karakteristik Bahasa Informal pada <i>Dataset</i>	49
Tabel 4.4 Perbandingan Tokenisasi	51
Tabel 4.5 Rata-rata Jumlah <i>Subword</i>	52
Tabel 4.6 Cakupan <i>Vocabulary</i> dan Tingkat <i>OOV</i>	54
Tabel 4.7 Statistik Distribusi Fragmentasi <i>Subword</i> per Kata	55
Tabel 4.8 Analisis <i>Vocabulary Overlap</i> antar <i>Tokenizer</i> BERT	56
Tabel 4.9 Statistik Panjang <i>Token</i> per <i>Tweet</i>	57
Tabel 4.10 Audit Ketahanan Tokenisasi pada Fenomena Bahasa Informal	59
Tabel 4.11 Audit <i>Fairness</i> Tokenisasi Antar Kelas Sentimen	61
Tabel 4.12 Lima Belas Contoh Kata <i>OOV</i> di <i>TF-IDF</i> (Hilang dari Representasi)	62
Tabel 4.13 Sepuluh Kata dengan Fragmentasi Paling Parah oleh mBERT	63
Tabel 4.14 Sepuluh Kata dengan Fragmentasi Paling Parah oleh IndoBERT	64
Tabel 4.15 Contoh Kata yang Dipecah mBERT tetapi Utuh oleh IndoBERT	65
Tabel 4.16 Contoh Kata yang Utuh oleh mBERT tetapi Dipecah oleh IndoBERT	66
Tabel 4.17 Tiga <i>Tweet</i> dengan Gap Tokenisasi mBERT vs IndoBERT Paling Besar	68
Tabel 4.18 Dua <i>Tweet</i> dengan Rasio <i>OOV</i> Tinggi pada <i>TF-IDF</i>	69
Tabel 4.19 Ringkasan Audit <i>Tokenizer</i> Tiga Pendekatan	70
Tabel 4.20 Hasil Pencarian <i>Hyperparameter</i> Tahap Pertama (1 <i>Epoch</i>)	71
Tabel 4.21 Perkembangan <i>Validation Loss</i> dan <i>Accuracy</i> Validasi per <i>Epoch</i>	72
Tabel 4.22 Perbandingan Performa mBERT dan IndoBERT pada <i>Test set</i>	73
Tabel 4.23 <i>Classification Report</i> Per Kelas Sentimen	75
Tabel 4.24 Sepuluh Konfigurasi <i>TF-IDF</i> + <i>SVM</i> Teratas dari <i>Grid search</i>	76
Tabel 4.25 Metrik Performa <i>TF-IDF</i> + <i>Linear SVM</i> pada <i>Test set</i>	78
Tabel 4.26 <i>Classification Report</i> <i>TF-IDF</i> + <i>Linear SVM</i> per Kelas	78
Tabel 4.27 Sepuluh Kata Paling Diskriminatif per Kelas Sentimen	79
Tabel 4.28 Perbandingan Performa Tiga Model pada <i>Test set</i> yang Identik	80
Tabel 4.29 Hasil Ablasi <i>Class weighting</i> pada Ketiga Model	81
Tabel 4.30 Train Loss, Train Accuracy, Val Loss, Val Accuracy, dan Gap Overfitting per Epoch	83
Tabel 4.31 Performa mBERT vs IndoBERT per Kategori Karakteristik <i>Tweet</i>	87
Tabel 4.32 <i>Accuracy</i> versus Selisih Fragmentasi (mBERT – IndoBERT)	88
Tabel 4.33 <i>Error Rate</i> per Kelas Sentimen	90
Tabel 4.34 Arah Kesalahan Klasifikasi mBERT	90
Tabel 4.35 Perbandingan Karakteristik Linguistik: Prediksi Benar vs Salah	91
Tabel 4.36 <i>CI Bootstrap</i> 95% untuk <i>F1-macro</i>	95

Tabel 4.37 Hasil Uji <i>McNemar</i> untuk Kontras Antar Model.....	97
Tabel 4.38 Hasil Uji t statistik untuk Kontras Antar Model	98



DAFTAR GAMBAR

	(halaman)
Gambar 2.1 Arsitektur <i>Transformers</i> (Vaswani dkk., 2017)	11
Gambar 2.2 Representasi <i>Input</i> BERT (Devlin dkk., 2019)	12
Gambar 2.3 Ilustrasi Tahapan <i>Pre-training</i> dan <i>Fine-tuning</i> (Devlin dkk., 2019) ..	13
Gambar 2.4 Proses <i>Masked Language Modelling</i> (MLM)	13
Gambar 2.5 Ilustrasi <i>fine-tuning</i> BERT untuk Klasifikasi (Devlin dkk., 2019) ...	14
Gambar 3.1 Diagram Alir Metode Penelitian	37
Gambar 4.1 Distribusi Label Sentimen pada <i>Dataset</i>	48
Gambar 4.2 Distribusi Panjang <i>Tweet</i> dan Sebarannya per Kelas Sentimen	49
Gambar 4.3 Persentase <i>Tweet</i> yang Memuat Karakteristik Bahasa Informal	50
Gambar 4.4 Rata-rata <i>Subword</i> per Kata pada mBERT dan IndoBERT per Kategori	53
Gambar 4.5 Perbandingan <i>Type coverage</i> dan <i>Token coverage</i> pada Tiga <i>Tokenizer</i>	55
Gambar 4.6 Distribusi Jumlah <i>Subword</i> per Kata pada mBERT dan IndoBERT ..	56
Gambar 4.7 Diagram Venn <i>Vocabulary</i> Overlap antara mBERT dan IndoBERT ..	57
Gambar 4.8 Distribusi Panjang Tokenisasi per <i>Tweet</i>	58
Gambar 4.9 Audit Ketahanan Tokenisasi per Kategori Linguistik Twitter Indonesia	60
Gambar 4.10 Audit <i>Fairness</i> Tokenisasi Antar Kelas Sentimen	61
Gambar 4.11 Perbandingan Metrik Klasifikasi mBERT dan IndoBERT pada <i>Test</i> <i>set</i>	74
Gambar 4.12 <i>Confusion matrix</i> mBERT (kiri) dan IndoBERT (kanan) pada <i>Test</i> <i>set</i>	75
Gambar 4.13 <i>Confusion matrix</i> <i>TF-IDF</i> + <i>Linear SVM</i> pada <i>Test set</i>	79
Gambar 4.14 Visualisasi Koefisien <i>SVM</i> untuk Fitur Diskriminatif Per Kelas	80
Gambar 4.15 Kurva <i>Train</i> vs <i>Val Loss</i> dan <i>Train</i> vs <i>Val Accuracy</i> untuk mBERT	84
Gambar 4.16 Kurva <i>Train</i> vs <i>Val Loss</i> dan <i>Train</i> vs <i>Val Accuracy</i> untuk IndoBERT	84
Gambar 4.17 <i>F1-macro</i> mBERT dan IndoBERT per Kategori Karakteristik <i>Tweet</i>	88
Gambar 4.18 <i>CI Bootstrap</i> mBERT, IndoBERT, dan <i>TF-IDF SVM</i>	96

DAFTAR LAMPIRAN

	(halaman)
Lampiran 1. <i>Preprocessed dataset</i>	110
Lampiran 2. Contoh Hasil <i>Token</i> 34 Kata Uji	113
Lampiran 3. Tokenisasi per- <i>tweet</i>	115
Lampiran 4. Hasil Ablasi mBERT, IndoBERT, dan <i>TF-IDF SVM</i>	118
Lampiran 5. Log Pelatihan mBERT dan IndoBERT	119

