

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan pesat platform media sosial berbasis video pendek telah menjadikan aplikasi TikTok sebagai salah satu layanan hiburan dan interaksi digital paling populer di Indonesia. Tingginya antusiasme pengguna secara langsung berdampak pada lonjakan volume interaksi di platform distribusi aplikasi, khususnya melalui kolom ulasan (*review*) pengguna di Google Play Store. Ulasan ini memuat informasi krusial berupa opini, keluhan, hingga laporan *bug* yang merupakan aset data berharga bagi pengembang aplikasi. Pendekatan *text mining* melalui Analisis Sentimen diperlukan untuk mengklasifikasikan polaritas opini ke dalam kategori positif, negatif, atau netral secara otomatis. Namun, seperti yang disoroti oleh Wirayudha (2025), tantangan utama dalam mengklasifikasikan ulasan dari Google Play Store adalah volume data yang masif serta distribusi kelas sentimen yang sering kali tidak seimbang (*imbalanced data*).

Selain masalah distribusi kelas, tantangan terbesar lainnya adalah tingginya tingkat *noise* pada data. Karakteristik ulasan pengguna aplikasi hiburan di Indonesia sangat didominasi oleh penggunaan bahasa tidak baku, *slang* (bahasa gaul), singkatan, serta salah ketik (*typo*). Penelitian yang dilakukan oleh Dwi et al. (2025) menunjukkan bahwa algoritma *machine learning* tradisional (seperti Naïve Bayes atau *Support Vector Machine*) yang menggunakan ekstraksi fitur berbasis frekuensi kata (TF-IDF) memiliki kelemahan mendasar. Algoritma tradisional tersebut kesulitan menangkap konteks kalimat secara utuh, sehingga rentan mengalami penurunan akurasi yang signifikan ketika berhadapan dengan *noise* teks ulasan Play Store yang tidak terstruktur.

Untuk mengatasi keterbatasan pemahaman konteks tersebut, pemodelan berbasis *Deep Learning* dengan arsitektur *Transformer* yang diperkenalkan oleh Devlin et al. (2019) melalui model BERT (*Bidirectional Encoder Representations from Transformers*) terbukti memberikan lonjakan performa secara revolusioner di bidang pemrosesan bahasa alami. Secara spesifik untuk bahasa Indonesia, Koto et al. (2020) mengembangkan *pre-trained language model* bernama IndoBERT yang dilatih menggunakan jutaan korpus teks lokal. IndoBERT memiliki kemampuan untuk memetakan representasi kata (*contextual word embedding*) secara dua arah. Kemampuan ini terbukti sangat tangguh, di mana penelitian Zahara et al. (2025) menegaskan bahwa IndoBERT secara efektif mampu menggali dan memahami makna semantik dari komentar TikTok yang penuh dengan bahasa *slang* dan informal yang tidak ada di dalam kamus baku.

Meskipun IndoBERT sangat tangguh dalam pemahaman semantik, ulasan pada aplikasi digital umumnya memiliki karakteristik teks yang relatif pendek dan langsung pada intinya. Untuk mengekstraksi fitur lokal yang menjadi penentu utama polaritas sentimen, seperti susunan *n-gram* atau frase spesifik, penggunaan *Convolutional Neural Network* (CNN) 1-Dimensi menjadi arsitektur pendamping yang sangat ideal. Studi yang dilakukan oleh (Imron, Setiawan, & Santoso, 2023) membuktikan bahwa integrasi antara metode representasi kata dari *IndoBERT Embedding* dengan arsitektur CNN terbukti sangat berhasil meningkatkan akurasi deteksi sentimen pada ulasan *e-commerce*. CNN beroperasi secara paralel melalui operasi konvolusi, yang memungkinkan model untuk menyeleksi fitur (*feature map*) paling krusial secara komputasi dengan kecepatan *training* yang jauh lebih baik dibandingkan pemodelan sekuensial tunggal.

Berdasarkan celah penelitian (*research gap*) dan argumentasi teknis di atas, integrasi pemahaman konteks bahasa gaul dari IndoBERT dengan kecepatan ekstraksi fitur lokal dari CNN diharapkan dapat menghasilkan model klasifikasi sentimen yang unggul. Oleh karena itu, penelitian ini akan

mengangkat judul "Analisis Sentimen Ulasan Aplikasi TikTok di Google Play Store Menggunakan Arsitektur Hibrid IndoBERT dan CNN", dengan tujuan menghasilkan model prediksi sentimen berskala bahasa Indonesia informal dengan tingkat akurasi yang tinggi.

1.2 Rumusan Masalah

Berdasarkan konteks tersebut, maka rumusan masalah dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana mengimplementasikan arsitektur hibrid IndoBERT dan Convolutional Neural Network (CNN) 1-Dimensi untuk melakukan analisis sentimen pada teks ulasan aplikasi TikTok berbahasa Indonesia di Google Play Store?
2. Bagaimana efektivitas tahapan prapemrosesan (*text preprocessing*) dalam menangani karakteristik data ulasan yang penuh *noise* berupa bahasa *slang*, singkatan, dan ketidakbakuan kata sebelum diekstraksi oleh model IndoBERT?

1.3 Batasan Masalah

1. Sumber Data: Data ulasan (*review*) yang digunakan hanya diambil dari halaman aplikasi TikTok yang terdapat di platform distribusi digital **Google Play Store**, dan tidak mencakup data dari platform lain seperti Apple App Store atau media sosial lainnya.
2. Karakteristik Teks: Ulasan yang ditarik dan dianalisis adalah ulasan yang menggunakan Bahasa Indonesia, termasuk variasi bahasa tidak baku, singkatan, dan *slang* (bahasa gaul) yang umum digunakan oleh pengguna. Ulasan yang sepenuhnya menggunakan bahasa asing akan dihapus pada tahap prapemrosesan.
3. Volume dan Rentang Data: Pengambilan data (*web scraping*) dibatasi pada 5.000 ulasan terbaru yang ditulis dalam rentang waktu Januari 2026 sampai

dengan Juni 2026, untuk memastikan relevansi opini pengguna terhadap kondisi aplikasi TikTok saat ini.

4. Kategorisasi Sentimen: Pelabelan data sentimen diklasifikasikan ke dalam 3 (tiga) kelas polaritas, yaitu Positif, Netral, dan Negatif. Pelabelan ini diinisialisasi berdasarkan *rating* bintang pada ulasan (Bintang 1-2 = Negatif, Bintang 3 = Netral, Bintang 4-5 = Positif).
5. Fokus Algoritma: Model analisis sentimen secara spesifik dibangun menggunakan arsitektur hibrid IndoBERT (sebagai *word embedding* / ekstraksi fitur semantik) dan Convolutional Neural Network (CNN) 1D (sebagai pengklasifikasi), tanpa melakukan perbandingan dengan arsitektur *deep learning* lainnya.
6. Metrik dan Perangkat Bantu: Evaluasi performa model diukur menggunakan *Confusion Matrix* yang meliputi nilai Akurasi, *Precision*, *Recall*, dan *F1-Score*. Implementasi pengkodean dilakukan menggunakan bahasa pemrograman Python di dalam lingkungan komputasi Google Colab.

1.4 Tujuan Penelitian

Menerapkan arsitektur hibrid IndoBERT dan Convolutional Neural Network (CNN) 1-Dimensi dalam membangun model analisis sentimen pada dataset ulasan aplikasi TikTok berbahasa Indonesia yang bersumber dari Google Play Store. Secara rinci, tujuan penelitian ini adalah:

1. Mengimplementasikan arsitektur hibrid IndoBERT dan Convolutional Neural Network (CNN) 1-Dimensi sebagai model klasifikasi untuk melakukan analisis sentimen pada teks ulasan aplikasi TikTok berbahasa Indonesia.
2. Melakukan tahapan prapemrosesan teks (*text preprocessing*) yang optimal untuk membersihkan data ulasan dari *noise*, bahasa *slang*, singkatan, dan ketidakbakuan tata bahasa, sehingga proses

pengekstraksian fitur semantik oleh IndoBERT dapat berjalan maksimal.

1.5 Manfaat Penelitian

1. Manfaat Teoritis (Akademis) : Memperkaya literatur dan kajian keilmuan di bidang *Natural Language Processing* (NLP) dan *Text Mining*, khususnya mengenai efektivitas penerapan arsitektur hibrid IndoBERT dan *Convolutional Neural Network* (CNN) pada teks berbahasa Indonesia yang bersifat informal.
2. Manfaat Praktis : Memberikan transparansi mengenai tren opini publik terhadap suatu layanan digital, sehingga masyarakat atau pengguna baru dapat mengetahui ringkasan persepsi kualitas aplikasi tanpa harus membaca ulasan secara manual.

1.6 Sistematika Penulisan

Pembuatan skripsi ini dilakukan dengan pembagian bab sebagai berikut:

Bab I: PENDAHULUAN

Pada bab ini membahas mengenai latar belakang pemilihan topik analisis sentimen ulasan TikTok, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

Bab II: TINJAUAN PUSTAKA

Bab ini menjelaskan teori mendasar yang mendukung penelitian, meliputi Analisis Sentimen, *Natural Language Processing* (NLP), *Deep Learning*, konsep *Word Embedding* kontekstual, arsitektur model bahasa IndoBERT, arsitektur 1D-CNN (*Convolutional Neural Network*), serta metrik Evaluasi Kinerja Model.

Bab III: METODOLOGI PENELITIAN

Bab ini berisi jenis dan fokus penelitian, sumber data sekunder dari Google Play Store, desain sistem hibrid yang dirancang, tahapan *text preprocessing* (*case folding, cleansing, normalization, dan stopword removal*), skenario

pembagian data latih dan data uji, perancangan model IndoBERT + 1D-CNN, hingga konfigurasi pelatihan model.

Bab IV: HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil eksekusi prapemrosesan teks, hasil pembagian data, implementasi arsitektur model pada PyTorch, rekaman proses pelatihan (*training log*), hasil evaluasi kinerja menggunakan *Confusion Matrix*, serta pembahasan menyeluruh atas performa model hibrid usulan berdasarkan hasil pengujian tersebut.

Bab V: KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan akhir dari seluruh rangkaian eksperimen dan pembahasan yang telah dilakukan, serta saran-saran akademis yang bermanfaat untuk pengembangan penelitian selanjutnya.

1.7 Penelitian Terkait

Tabel 1.7 Penelitian Terkait

Penulis dan Tahun	Judul	Hasil / Pencapaian
Dwi et al.(2025)	Analisis Sentimen Ulasan Aplikasi TikTok Menggunakan Convolutional Neural Network (<i>Jurnal RIGGS</i>)	Melakukan klasifikasi ulasan TikTok di Google Play Store dan membuktikan bahwa arsitektur 1D-CNN sangat efektif mengekstraksi fitur lokal teks pendek, mencapai akurasi sebesar 98,95% mengalahkan algoritma tradisional.
Zahara et al.(2025)	Analisis Sentimen Program Makan Siang Gratis pada TikTok dengan	Menunjukkan bahwa <i>pre-trained model</i> IndoBERT memiliki keunggulan mutlak dan sangat tangguh dalam membaca struktur

	Pendekatan NLP Berbasis IndoBERT (<i>Jurnal Infomedia</i>)	semantik bahasa <i>slang</i> (informal) di kolom komentar pengguna TikTok.
Pratama et al. (2024)	Perbandingan Metode IndoBERT Dengan CNN Untuk Analisis Sentimen Berbasis Aspek Terhadap Ulasan Pelanggan (<i>J-PTIHK Universitas Brawijaya</i>)	Memanfaatkan <i>embedding</i> IndoBERT dan mengombinasikannya dengan proses konvolusi CNN untuk menangkap polaritas sentimen pada ulasan <i>platform</i> aplikasi digital, menghasilkan nilai metrik Presisi dan <i>Recall</i> yang stabil.
Imron et al.(2023)	Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN (<i>Jurnal INSYST</i>)	Membuktikan bahwa integrasi ekstraksi vektor kata dari representasi <i>IndoBERT Embedding</i> yang dilewatkan ke dalam arsitektur CNN berhasil mempercepat komputasi pelatihan model secara signifikan melalui <i>parallel processing</i> .
Mursidah et al.(2025)	Klasifikasi Sentimen Google Play Store Aplikasi ChatGPT	Membuktikan bahwa penerapan metode IndoBERT pada ulasan Google Play Store sangat tangguh menghadapi teks tidak beraturan, dan berhasil

	Berbahasa Indonesia Berbasis IndoBERT (<i>Jurnal Minfo Polgan</i>)	mendistribusikan tiga kelas sentimen dengan nilai keseimbangan <i>F1-Score</i> yang sangat tinggi (88%).
Alfinsyah and Hartato (2025)	Evaluating the Impact of Random Over Sampling on IndoBERT Performance for Indonesian Sentiment Analysis (<i>Journal of Applied Informatics and Computing</i>)	Menyelesaikan masalah umum pelabelan <i>rating</i> bintang ulasan Google Play Store, yaitu ketidakseimbangan data (<i>imbalanced data</i>). Penelitian ini membuktikan bahwa penyeimbangan kelas secara signifikan meningkatkan performa akurasi IndoBERT pada ulasan kelas minoritas.
Tarwoto et al.(2025)	Analisis Sentimen Ulasan Aplikasi Mobile JKN di Google PlayStore Menggunakan IndoBERT (<i>Jurnal JTIK</i>)	Mengklasifikasikan kepuasan layanan dengan memanfaatkan vektor <i>contextual embedding</i> bawaan dari model IndoBERT, yang terbukti mampu memahami pola keluhan bahasa Indonesia yang tidak ada di dalam kamus EYD.
Ramadhan et al. (2025)	Analisis Perbandingan Model CNN dan IndoBERT	Melakukan uji perbandingan pada kinerja <i>filter</i> CNN dengan <i>attention mechanism</i> IndoBERT. Penelitian menyoroti bahwa

	Dalam Sentimen Berita Politik Indonesia (<i>Jurnal SENATIB</i>)	kombinasi keunggulan ekstraksi fitur lokal (CNN) dan pemahaman makna (<i>Transformer</i>) merupakan pendekatan ideal untuk NLP Indonesia.
Sumartha (2025)	Implementasi IndoBERT Untuk Analisis Sentimen Opini Publik Terhadap Kebijakan Kenaikan UKT (<i>Jurnal Informatika dan Teknik Elektro Terapan</i>)	Mendeteksi sentimen opini netizen dengan presisi dan sensitivitas tinggi, serta menekankan pentingnya tahap pembersihan data teks (<i>text preprocessing</i>) sebelum memasukkannya ke dalam <i>tokenizer</i> IndoBERT.
Koto et al. (2020)	IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP (<i>Proceedings of COLING</i>)	Merupakan jurnal pencipta (kertas kerja fundamental) dari IndoBERT yang secara matematis membuktikan bahwa arsitektur <i>bidirectional attention</i> jauh mengungguli performa pembobotan masa lalu seperti TF-IDF maupun Word2Vec.

