Teknik: Jurnal Ilmu Teknik dan Informatika
Volume 6 Nomor 2 Oktober 2026
e-ISSN: 2798-2513; p-ISSN: 2808-8751, Hal *27-40*
DOI: https://doi.org/10.51903/teknik.v6i2.1255
Available online at: https://journal.stiestekom.ac.id/index.php/TEKNIK

Rancang Bangun Aplikasi Web Identifikasi Teks Akademik dan Non-Akademik Berbasis IndoBERT sebagai Implementasi Pendekatan Natural Language Processing

Muhammad Tsabat Muhyiyuddin[1], Muhammad Zikry Sahendra Chan[2], Elga Sari Tanjung[3], Bicanro Gebriyan Panjaitan[4], Anggia Puteri[5]
[1-5]Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, Sumatera Utara, Indonesia

Korespondensi penulis: mtsabat.4243250002@mhs.unimed.ac.id

Abstract*. The ability to automatically distinguish academic from non-academic texts is an important need in higher education environments. This study aims to design and build a web application that implements the IndoBERT model to classify Indonesian texts into academic or non-academic categories with interpretable output accompanied by classification reasoning. The system is built end-to-end using the PyTorch and HuggingFace Transformers stack for modeling, FastAPI as the backend, and vanilla HTML/CSS/JavaScript for the user interface. The dataset was manually compiled with 209 samples (108 academic and 101 non-academic texts). Model fine-tuning was performed on Google Colab using a Tesla T4 GPU with a learning rate of 2×10^{-5}, batch size of 16, and eval_loss-based early stopping. The system is also equipped with a zero-shot fallback module so the application remains operational without training. On the internal test set (21 samples), the model achieved 100% accuracy with an F1-score of 1.0000 and an inference speed of 404 samples/second. This research is positioned as a proof-of-concept emphasizing the completeness of the transformer-based NLP pipeline and the interpretability of classification results, rather than a claim of state-of-the-art performance. Discussion of limitations and directions for further development are also presented.*
Keywords*: FastAPI; IndoBERT; Natural Language Processing; Text Classification; Web Application.*

Abstrak. Kemampuan membedakan teks akademik dan teks non-akademik secara otomatis merupakan kebutuhan penting di lingkungan perguruan tinggi. Penelitian ini bertujuan merancang dan membangun aplikasi web yang mengimplementasikan model IndoBERT untuk mengklasifikasikan teks berbahasa Indonesia ke dalam kategori akademik atau non-akademik, dengan output yang interpretable (disertai alasan klasifikasi). Sistem dibangun secara end-to-end menggunakan stack PyTorch dan HuggingFace Transformers untuk pemodelan, FastAPI sebagai backend, serta HTML/CSS/JavaScript vanilla untuk antarmuka pengguna. Dataset disusun manual sebanyak 209 sampel dengan komposisi 108 teks akademik dan 101 teks non-akademik. Fine-tuning model dilakukan pada Google Colab menggunakan GPU Tesla T4 dengan konfigurasi learning rate 2×10^{-5}, batch size 16, dan early stopping berbasis eval_loss. Sistem juga dilengkapi dengan modul fallback zero-shot sehingga aplikasi tetap dapat dijalankan tanpa proses training. Pada test set internal (21 sampel), model mencapai akurasi 100% dengan F1-score 1,0000 dan kecepatan inferensi 404 sampel/detik. Penelitian ini diposisikan sebagai proof-of-concept yang menekankan kelengkapan pipeline NLP berbasis transformer dan interpretabilitas hasil klasifikasi, bukan klaim performa state-of-the-art. Pembahasan limitasi serta arah pengembangan lanjutan turut disampaikan dalam artikel ini.
Kata kunci: Aplikasi Web; FastAPI; IndoBERT; Klasifikasi Teks; Natural Language Processing.

1. LATAR BELAKANG

Perkembangan teknologi informasi yang pesat telah menghasilkan ledakan volume teks digital berbahasa Indonesia, mulai dari jurnal ilmiah, skripsi, dan laporan penelitian, hingga artikel blog, berita daring, dan unggahan media sosial [1]. Di lingkungan

Diterima: 01 Juni 2026; Direvisi: 16 Juli 2026; Disetujui: 02 September 2026; Tersedia Online: 19 September 2026; Diterbitkan: 03 Oktober 2026.

perguruan tinggi, kemampuan membedakan teks akademik dari teks non-akademik merupakan kompetensi dasar yang penting bagi mahasiswa maupun tenaga pengajar.

Teks akademik dicirikan oleh penggunaan bahasa formal, struktur yang sistematis, serta dukungan referensi ilmiah [2]. Ciri-ciri tersebut mencakup sifat objektif, penggunaan kalimat efektif, berlandaskan fakta dan teori yang terverifikasi, serta kepatuhan terhadap konvensi penulisan ilmiah seperti sistem sitasi dan struktur baku [3]. Sebaliknya, teks non-akademik cenderung subjektif, menggunakan bahasa informal atau kolokial, dan tidak terikat pada metodologi atau referensi ilmiah [4].

Identifikasi jenis teks secara manual membutuhkan waktu dan keahlian yang tidak sedikit. Otomatisasi proses ini melalui pendekatan Natural Language Processing (NLP) menjadi solusi yang relevan. Salah satu kemajuan terpenting dalam NLP modern adalah arsitektur Transformer dan model BERT (Bidirectional Encoder Representations from Transformers) yang diperkenalkan oleh Devlin dkk. [5]. BERT memungkinkan pemodelan representasi kontekstual teks secara bidirectional, sehingga menghasilkan performa state-of-the-art pada berbagai tugas klasifikasi teks.

Untuk konteks bahasa Indonesia, Wilie dkk. [6] mengembangkan IndoBERT, sebuah model BERT yang dilatih pada korpus berbahasa Indonesia (Indo4B) sebagai bagian dari benchmark IndoNLU. IndoBERT telah terbukti efektif diterapkan pada berbagai tugas klasifikasi teks bahasa Indonesia, mulai dari analisis sentimen [7], klasifikasi emosi [8], hingga deteksi berita hoaks [9][10]. Meskipun demikian, pemanfaatan IndoBERT untuk identifikasi teks akademik versus non-akademik berbahasa Indonesia masih jarang ditemukan, khususnya dalam bentuk aplikasi web yang dapat langsung digunakan oleh pengguna akhir.

Penelitian ini bertujuan untuk merancang dan membangun aplikasi web end-to-end yang mengintegrasikan model IndoBERT untuk identifikasi teks akademik dan non-akademik berbahasa Indonesia. Sistem dirancang dengan tiga karakteristik utama: (1) pipeline NLP lengkap mulai dari penyusunan dataset, fine-tuning model, hingga deployment web; (2) mekanisme fallback dua mode (fine-tuned dan zero-shot) sehingga aplikasi tetap dapat dijalankan walaupun belum dilakukan training; dan (3) output klasifikasi yang interpretable melalui kombinasi prediksi model transformer dengan ekstraksi fitur linguistik. Penelitian ini diposisikan sebagai proof-of-concept yang

menekankan kelengkapan pipeline dan interpretabilitas, bukan klaim performa state-of-the-art.

2. KAJIAN TEORITIS

Teks Akademik dan Non-Akademik

Teks akademik adalah tulisan yang membahas ilmu pengetahuan, disusun secara sistematis dengan bahasa formal serta didukung referensi ilmiah [2]. Karakteristik utamanya meliputi penggunaan istilah teknis, kalimat kompleks, struktur sitasi, dan register formal-baku. Sebaliknya, teks non-akademik dalam konteks penelitian ini didominasi gaya bahasa kolokial yang lazim ditemukan di media sosial dan percakapan sehari-hari, dengan marker khas seperti penggunaan kata "gue", "banget", "sih", dan "deh" [3][4].

BERT dan IndoBERT

BERT [5] adalah model bahasa berbasis arsitektur Transformer yang dilatih dengan dua objektif utama: Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP). Pendekatan bidirectional ini memungkinkan BERT memahami konteks dari kedua arah kalimat secara simultan. Untuk konteks bahasa Indonesia, IndoBERT [6] dilatih pada korpus Indo4B yang dikumpulkan dari berbagai sumber publik seperti media sosial, blog, berita, dan situs web. Penelitian ini menggunakan varian indobenchmark/indobert-base-p1 dengan 12 layer encoder, hidden size 768, 12 attention head, dan total parameter sekitar 110 juta.

Fine-Tuning untuk Klasifikasi Teks

Fine-tuning adalah proses adaptasi model pre-trained ke tugas spesifik dengan dataset yang relatif kecil. Pada klasifikasi teks, fine-tuning BERT biasanya dilakukan dengan menambahkan lapisan klasifikasi linear di atas representasi token [CLS] dan melakukan training end-to-end dengan loss cross-entropy [5]. Berbagai penelitian telah menunjukkan efektivitas pendekatan ini untuk tugas-tugas klasifikasi teks bahasa Indonesia, antara lain klasifikasi emosi [8], analisis sentimen [7], dan deteksi berita hoaks menggunakan IndoBERT fine-tuning [9][10].

FastAPI sebagai Backend Aplikasi NLP

FastAPI [11] adalah framework web modern berbasis Python yang dirancang untuk membangun API dengan performa tinggi melalui dukungan asynchronous (ASGI), type hints Pydantic, dan dokumentasi otomatis melalui Swagger UI. Framework ini menjadi

pilihan populer untuk deployment model machine learning karena kemudahan integrasi dengan stack ML Python serta performa yang setara dengan Node.js dan Go [11].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan penelitian dan pengembangan dengan model waterfall yang mencakup empat tahapan: (1) analisis kebutuhan, (2) perancangan sistem, (3) implementasi, dan (4) pengujian. Penelitian ini diposisikan sebagai proof-of-concept dengan fokus pada kelengkapan pipeline NLP berbasis transformer.

Stack Teknologi

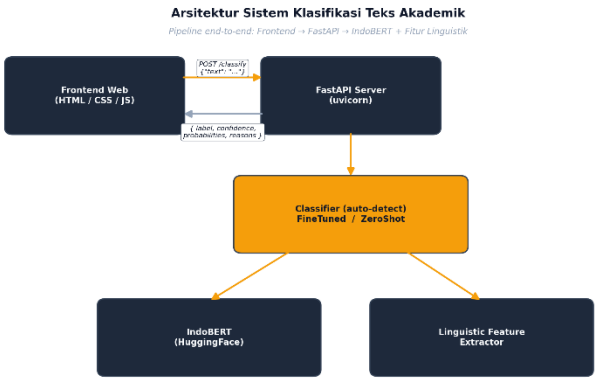
Sistem dibangun menggunakan kombinasi teknologi sebagaimana disajikan pada Tabel 1 berikut.

Komponen	Teknologi
Pre-trained Model	indobenchmark/indobert-base-p1
Framework ML	PyTorch + HuggingFace Transformers 4.44.2
Backend	FastAPI 0.115 + Uvicorn
Frontend	HTML + CSS + JavaScript vanilla
Training Environment	Google Colab (GPU Tesla T4, 16 GB VRAM)
Bahasa Pemrograman	Python 3.11
Splitting & Metrik	scikit-learn 1.5.1

Tabel 1. Stack teknologi sistem

Arsitektur Sistem

Arsitektur sistem dirancang sebagai aplikasi client-server. Frontend mengirimkan teks input melalui HTTP POST ke endpoint /classify pada FastAPI server. Server kemudian mendeteksi secara otomatis apakah model fine-tuned tersedia. Jika tersedia, FineTunedClassifier akan digunakan; jika tidak, sistem akan jatuh kembali ke ZeroShotClassifier. Hasil klasifikasi dikembalikan dalam format JSON yang berisi label, confidence, distribusi probabilitas, fitur linguistik, dan alasan klasifikasi.



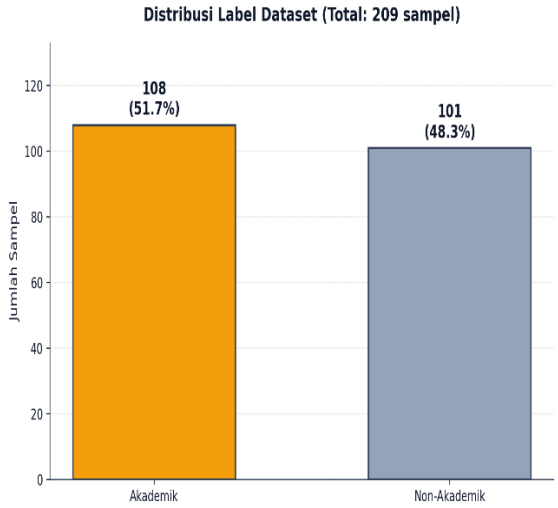
Gambar 1. Diagram arsitektur sistem klasifikasi

Dataset

Dataset disusun manual oleh peneliti dengan format CSV berisi dua kolom (text, label). Total dataset berjumlah 209 baris dengan komposisi 108 sampel akademik dan 101 sampel non-akademik (51,7% vs 48,3%) sehingga distribusi mendekati seimbang. Dataset mencakup variasi domain seperti manajemen, pendidikan, kesehatan, teknik, sosial, dan lingkungan, serta variasi panjang kalimat.

ID	Teks	Label
1	Penelitian ini bertujuan untuk menganalisis pengaruh kualitas pelayanan terhadap kepuasan pelanggan pada industri perbankan syariah di Indonesia.	akademik
2	Berdasarkan hasil uji regresi linier berganda, ditemukan bahwa terdapat pengaruh yang signifikan antara variabel motivasi kerja terhadap kinerja karyawan dengan nilai p sebesar 0,003.	akademik
3	Metode yang digunakan dalam penelitian ini adalah pendekatan kuantitatif dengan teknik pengumpulan data melalui kuesioner yang disebarkan kepada 150 responden.	akademik
4	Studi literatur menunjukkan bahwa teori perilaku konsumen yang dikemukakan oleh Kotler relevan untuk menjelaskan fenomena pembelian impulsif pada generasi milenial.	akademik
5	Analisis data dilakukan dengan menggunakan perangkat lunak SPSS versi 25 untuk menguji validitas dan reliabilitas instrumen penelitian.	akademik
6	Kesimpulan dari penelitian ini menunjukkan adanya kontribusi teoretis dan praktis terhadap pengembangan ilmu manajemen sumber daya manusia.	akademik
7	Hipotesis yang diajukan dalam penelitian ini berhasil diterima berdasarkan hasil pengujian statistik dengan tingkat signifikansi 5 persen.	akademik
8	Populasi dalam penelitian ini adalah seluruh mahasiswa aktif Fakultas Ekonomi angkatan 2021 hingga 2023 dengan total 1.250 orang.	akademik
9	Teknik pengambilan sampel menggunakan purposive sampling dengan kriteria responden telah menggunakan aplikasi e-commerce minimal enam bulan terakhir.	akademik
10	Variabel independen dalam penelitian ini meliputi kualitas produk, harga, dan promosi, sedangkan variabel dependennya adalah keputusan pembelian.	akademik

Gambar 2. Sampel CSV dataset



Gambar 3. Output distribusi label menggunakan value_counts()

Karakteristik Linguistik Pembeda

Untuk mendukung interpretabilitas hasil, sistem mengekstraksi sejumlah fitur linguistik dari teks input. Karakteristik pembeda antara dua kelas dirangkum pada Tabel 2.

Aspek	Akademik	Non-Akademik
Register	Formal, baku	Informal, kolokial
Topik dominan	Penelitian, metodologi, analisis statistik	Curhat, gosip, belanja, hiburan
Marker khas	"penelitian ini", "berdasarkan hasil", "signifikan"	"gue", "banget", "gak", "yaudah"
Panjang rata-rata	15–35 kata/kalimat	12–25 kata/kalimat
Struktur kalimat	Kompleks, dependen	Sederhana, percakapan

Tabel 2. Karakteristik linguistik pembeda teks akademik vs non-akademik

Preprocessing

Preprocessing dilakukan dalam beberapa tahap: (1) strip whitespace dan validasi label; (2) lowercase pada label untuk konsistensi; (3) tokenisasi menggunakan tokenizer bawaan IndoBERT berbasis WordPiece; (4) penetapan max sequence length 128 token; serta (5) padding dinamis melalui DataCollatorWithPadding untuk efisiensi memori.

Pembagian Data

Data dibagi menggunakan stratified random split untuk mempertahankan proporsi label, dengan random seed 42 agar dapat direproduksi. Proporsi pembagian: 80% train, 10% validation, dan 10% test, dengan rincian seperti pada Tabel 3.

Subset	Total	Akademik	Non-Akademik
Train	167	86	81
Validation	21	11	10
Test	21	11	10

Tabel 3. Pembagian dataset (stratified, seed=42)

Arsitektur Model dan Hyperparameter

Model yang digunakan adalah IndoBERT (indobenchmark/indobert-base-p1) dengan classification head berupa lapisan linear (768 → 2) dan fungsi aktivasi Softmax. Loss function yang digunakan adalah Cross-Entropy. Hyperparameter training dirangkum pada Tabel 4.

Parameter	Nilai
Learning rate	2×10^{-5}
Optimizer	AdamW
Weight decay	0,01
Warmup ratio	0,1 (10% total step)
Batch size (train)	16
Batch size (eval)	32
Epoch maksimum	6
Early stopping patience	2 (berdasarkan eval_loss)
Max sequence length	128 token
Mixed precision	fp16 (saat GPU tersedia)
Seed	42

Tabel 4. Hyperparameter training fine-tuning IndoBERT

Lingkungan Eksekusi

Proses training dilakukan di Google Colab dengan GPU Tesla T4. Inferensi (deployment) dilakukan pada lingkungan lokal Windows 11 berbasis CPU. Durasi training aktual berkisar 1–2 menit (6 epoch dengan early stopping aktif) — jauh lebih cepat dibandingkan training pada CPU yang dapat memakan waktu sekitar 15 menit.

4. HASIL DAN PEMBAHASAN

Hasil Training

Proses fine-tuning IndoBERT berhasil dilakukan pada Google Colab dengan GPU Tesla T4. Early stopping aktif berdasarkan metrik eval_loss dengan patience 2 epoch.

[Gambar 4. Progress training (tabel epoch dan loss) di Google Colab]

Evaluasi pada Test Set

Evaluasi model dilakukan pada test set internal sebanyak 21 sampel. Hasil metrik evaluasi dirangkum pada Tabel 5.

Metrik	Nilai
Accuracy	1,0000 (100%)
Precision (macro)	1,0000
Recall (macro)	1,0000
F1-score (macro)	1,0000
Test loss	0,0006
Inference speed	404 sampel/detik

Tabel 5. Metrik evaluasi model pada test set

Penting untuk ditekankan bahwa akurasi 100% diperoleh pada test set yang berasal dari distribusi yang sama dengan training (split internal). Hasil ini tidak menunjukkan generalisasi sempurna ke teks dunia nyata, melainkan menggambarkan kemampuan model dalam mempelajari pola permukaan dalam dataset yang relatif kecil dan terkurasi. Pembahasan limitasi lebih lanjut disampaikan pada Sub-bab 4.6.

```
Classification report:
              precision    recall  f1-score   support

non-akademik     1.0000    1.0000    1.0000        10
    akademik     1.0000    1.0000    1.0000        11

    accuracy                         1.0000        21
   macro avg     1.0000    1.0000    1.0000        21
weighted avg     1.0000    1.0000    1.0000        21
```

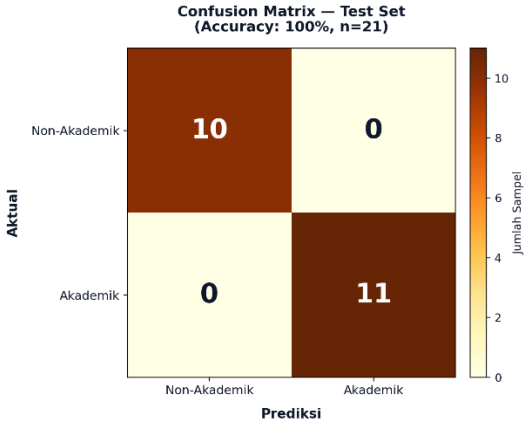
Gambar 5. Classification report hasil evaluasi di Google Colab

Confusion Matrix

Confusion matrix pada test set menunjukkan tidak ada kesalahan klasifikasi pada distribusi internal, sebagaimana disajikan pada Tabel 6.

	Pred: Non-Akademik	Pred: Akademik
Aktual: Non-Akademik	10	0
Aktual: Akademik	0	11

Tabel 6. Confusion matrix pada test set



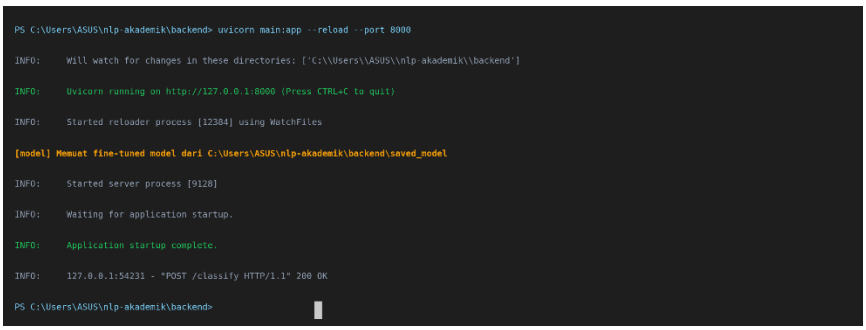
Gambar 6. Visualisasi confusion matrix (matplotlib)

4.4 Implementasi Aplikasi Web

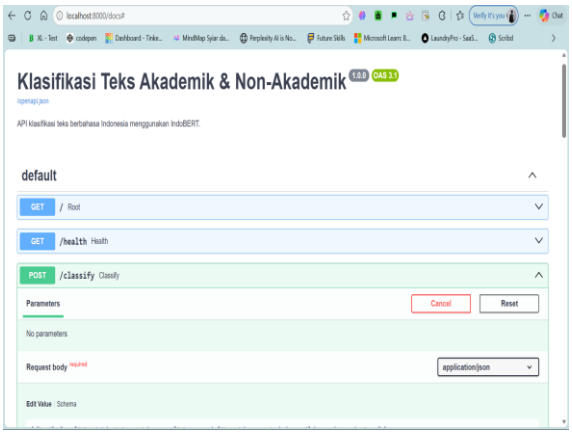
Aplikasi web berhasil dibangun dengan struktur folder modular yang memisahkan komponen backend, dataset, frontend, dan notebook training. Server FastAPI dapat diakses melalui Uvicorn pada localhost:8000, dengan dokumentasi API otomatis tersedia melalui Swagger UI pada endpoint /docs.

```
nlp-akademik/
├── backend/
│   ├── main.py
│   ├── model.py
ZeroShotClassifier
│   ├── train.py
│   ├── saved_model/
tokenizer, metrics.json)
│   └── requirements.txt
├── dataset/
│   └── dataset.csv
├── train_colab.ipynb
├── frontend/
│   ├── index.html
│   ├── style.css
│   └── script.js
└── README.md
```

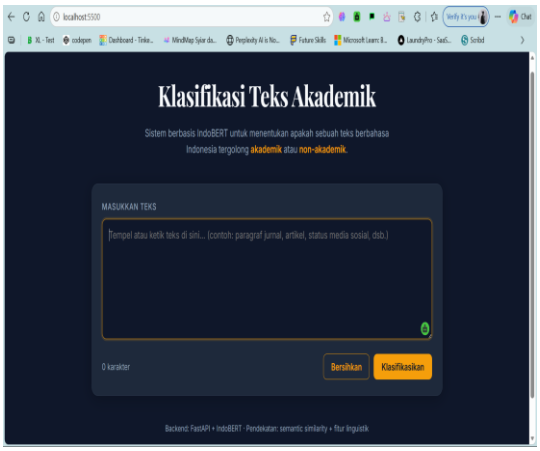
Gambar 7. Struktur folder project pada VSCode file explorer



Gambar 8. Log Uvicorn menampilkan pesan [model] Memuat fine-tuned model



Gambar 9. Tampilan Swagger UI pada http://localhost:8000/docs



Gambar 10. Tampilan awal antarmuka web (form input kosong)

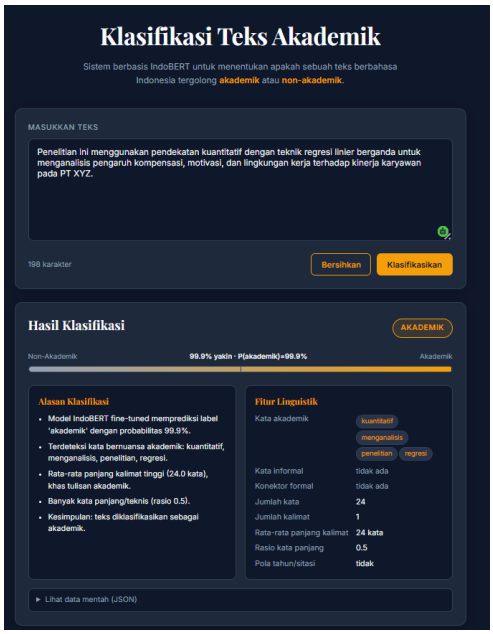
Contoh Output Klasifikasi

Sistem dirancang untuk menghasilkan output yang interpretable. Selain label dan confidence, output juga menyertakan fitur linguistik yang diekstraksi dari teks serta daftar alasan klasifikasi. Contoh hasil klasifikasi pada beberapa teks uji disajikan pada Tabel 7.

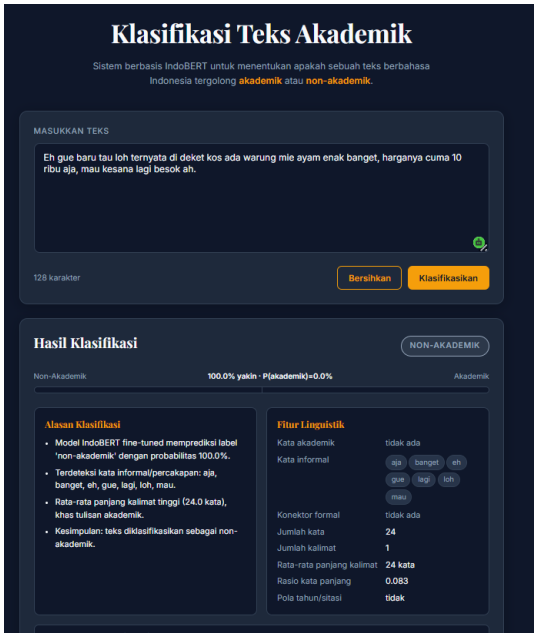
Input Teks	Prediksi	Catatan
"Penelitian ini bertujuan menganalisis pengaruh literasi keuangan terhadap perilaku menabung mahasiswa."	Akademik	Benar (confidence tinggi)
"Hasil uji regresi menunjukkan bahwa variabel motivasi berpengaruh signifikan terhadap kinerja."	Akademik	Benar
"Eh tau gak sih, kemarin gue jalan-jalan ke mall sama temen, terus makan bakso enak banget!"	Non-Akademik	Benar

Input Teks	Prediksi	Catatan
"Cuaca lagi panas banget nih, mending rebahan aja di rumah sambil main game."	Non-Akademik	Benar
"Refleksi hari ini mengajarkan bahwa konsistensi adalah kunci dari pencapaian jangka panjang."	Akademik	Salah (seharusnya non-akademik — caption motivasi)

Tabel 7. Contoh hasil klasifikasi pada teks uji



Gambar 11. Hasil klasifikasi teks akademik di antarmuka web



Gambar 12. Hasil klasifikasi teks non-akademik di antarmuka web

Selain label dan confidence, sistem menghasilkan output JSON yang berisi distribusi probabilitas, daftar kata akademik dan kata informal yang terdeteksi, konektor formal, jumlah kata, jumlah kalimat, rasio kata panjang, dan daftar reasons yang dapat dibaca pengguna. Contoh format output JSON dari endpoint /classify:

{ "mode": "fine-tuned", "label": "akademik", "confidence": 0.9876, "probabilities": { "akademik": 0.9876, "non_akademik": 0.0124 }, "features": { "kata_akademik": ["penelitian", "metode", "analisis"], "konektor_formal": ["berdasarkan"], "rata_panjang_kalimat": 16.0 }, "reasons": ["Model IndoBERT fine-tuned memprediksi label akademik dengan probabilitas 98.8%.", "Terdeteksi kata bernuansa akademik: penelitian, metode, analisis.", "Menggunakan konektor formal: berdasarkan."] }

Listing 1. Contoh format output JSON dari endpoint /classify

Pembahasan dan Limitasi

Hasil akurasi 100% pada test set merupakan indikator yang menggembirakan, tetapi harus dibaca dengan hati-hati dan tidak boleh dibesar-besarkan. Beberapa limitasi penting penelitian ini adalah sebagai berikut:

A. Ukuran dataset yang kecil (209 sampel) berpotensi menyebabkan overfitting. Model cenderung "menghafal" pola permukaan dataset daripada memahami konsep mendalam tentang teks akademik.
B. Distribusi non-akademik dalam dataset didominasi gaya kolokial seperti yang lazim ditemukan di Twitter dan WhatsApp. Dataset belum mencakup non-akademik formal seperti caption motivasi, berita, opini editorial, atau narasi naratif.
C. Sumber dataset bersifat tunggal (disusun manual oleh peneliti), sehingga berpotensi mengandung bias subjektif.
D. Test set hanya berukuran 21 baris dengan distribusi yang sama dengan training, sehingga akurasi 100% tidak menggambarkan kemampuan generalisasi pada teks dunia nyata.
E. Pengujian dengan teks edge case seperti kalimat refleksi motivasi formal (lihat baris terakhir Tabel 7) menunjukkan bahwa model cenderung mengasosiasikan register formal dengan label akademik, bukan substansi keilmiahan. Hal ini merupakan artefak dataset, bukan kelemahan model.
F. Model belum diuji pada teks panjang; input lebih dari 128 token akan dipotong.
G. Dengan limitasi tersebut, sistem yang dibangun lebih tepat diposisikan sebagai proof-of-concept untuk demonstrasi pipeline NLP berbasis transformer yang

interpretable. Sistem belum cocok untuk lingkungan produksi tanpa perluasan dataset signifikan dan redefinisi label.

5. KESIMPULAN DAN SARAN

Penelitian ini telah berhasil merancang dan membangun aplikasi web end-to-end untuk identifikasi teks akademik dan non-akademik berbahasa Indonesia menggunakan model IndoBERT. Sistem terdiri dari pipeline lengkap mulai dari penyusunan dataset (209 sampel), fine-tuning model di Google Colab, hingga deployment menggunakan FastAPI dan frontend berbasis HTML/CSS/JavaScript vanilla. Mekanisme fallback dua mode (fine-tuned dan zero-shot) memastikan aplikasi tetap dapat dijalankan walaupun belum ada model hasil training. Output klasifikasi yang interpretable dihasilkan melalui kombinasi prediksi IndoBERT dengan ekstraksi fitur linguistik. Pada test set internal, model mencapai akurasi 100% dengan F1-score 1,0000 dan kecepatan inferensi 404 sampel/detik — meski hasil ini terbatas pada distribusi internal dan tidak menggambarkan generalisasi ke teks dunia nyata.

Sebagai saran pengembangan lanjutan, penelitian berikutnya dapat: (1) memperluas dataset menjadi minimal 2.000+ sampel per kelas dengan sumber beragam (jurnal Garuda, repositori skripsi, media sosial, berita); (2) menambahkan kategori "non-akademik formal" untuk mengatasi bias register; (3) meredefinisi label menjadi "berbasis penelitian empiris vs lainnya" agar lebih kuat secara konseptual; (4) menerapkan cross-validation k-fold sebagai pengganti single split untuk evaluasi yang lebih robust; (5) membandingkan dengan IndoBERT-large, XLM-R, mBERT, atau IndoRoBERTa; serta (6) mengeksplorasi multi-label classification untuk teks campuran.

DAFTAR REFERENSI

Tsirmpas, D., Gkionis, I., Papadopoulos, G. T., & Mademlis, I. (2024). Neural Natural Language Processing for Long Texts: A Survey on Classification and Summarization. Engineering Applications of Artificial Intelligence, 133, 108231. https://doi.org/10.1016/j.engappai.2024.108231

Halawa, Y., Pardede, N. F., Tambusai, B. A. D., Damanik, F. S., & Purnama, I. (2025). Memahami Teks Akademik: Struktur, Karakteristik, dan Peran Dalam Pengembangan Ilmu. Jurnal TIPS: Jurnal Riset, Pendidikan dan Ilmu Sosial, 3(1).

Irsyad, F. R., Nainggolan, N. P., Gultom, R. H., Sari, A., & Lubis, F. (2023). Struktur Dan Pengaruh Teks Akademik Dan Non-Akademik Dalam Meningkatkan Pemahaman Mahasiswa Pada Proses Pembelajaran. Jurnal Bima: Pusat Publikasi Ilmu Pendidikan Bahasa dan Sastra, 1(4), 310–325. https://doi.org/10.61132/bima.v1i4.307

Manik, A., Thesia, D. P., & Tarigan, E. P. B. (2024). Kajian Penulisan Teks Akademik Bagi Mahasiswa Terhadap Peningkatan Keterampilan Menulis. Bhinneka: Jurnal Bintang Pendidikan dan Bahasa, 2(3), 115–122. https://doi.org/10.59024/bhinneka.v2i3.845

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://aclanthology.org/N19-1423/

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. Proceedings of AACL-IJCNLP 2020. https://aclanthology.org/2020.aacl-main.85/

Dalimunthe, A. H., Ula, M., & Meiyanti, R. (2024). Unjuk Kerja Algoritma Support Vector Machine (SVM) dan Naïve Bayes Dalam Pengklasifikasian Berita Hoaks Pada Twitter. Jurnal Elektronika dan Teknologi Informasi, 5(2), 11–22. https://doi.org/10.5201/jet.v5i2.400

Rizky, M., dkk. (2024). Klasifikasi Emosi Pada Teks Bahasa Indonesia Menggunakan IndoBERT. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika).

Kunaefi, A., dkk. (2025). Klasifikasi Berita Hoaks Bahasa Indonesia Menggunakan IndoBERT Fine-Tuning dengan Pendekatan Focal Loss pada Data Tidak Seimbang. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika).

Wijayakusuma, I. G. N. L., dkk. (2025). Perbandingan Kinerja IndoBERT dan MBERT untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia. JST: Jurnal Sains dan Teknologi, 14(1), 114–123. https://ejournal.undiksha.ac.id/index.php/JST/article/view/92126

Ramirez, S. (2018–2024). FastAPI Framework Documentation. https://fastapi.tiangolo.com/

Septiani, D., & Isabela, I. (2022). Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam Temu Kembali Informasi pada Dokumen Teks. Sistem dan Teknologi Informasi Indonesia (SINTESIA), 1(2), 81–88.

Dwifebri, P., & Suhardianto. (2023). Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet "Mental Health". DECODE: Jurnal Pendidikan Teknologi Informasi, 3(2), 161–171.

Rahmadhani, F., dkk. (2023). Analisis Pengaruh Penerapan Stopword Removal. JUARA: Jurnal Aplikasi dan Riset Informatika, 2(1).

Abdillah, G., dkk. (2021). Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification. Jurnal Teknik Informatika, 6(2).