Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer
Volume. 6 Nomor. 2 Juli 2026
e-ISSN: 2827-7953; p-ISSN: 2827-8135; Hal 208-221
DOI: https://doi.org/10.55606/juisik.v6i2.2411
Tersedia: https://journal.sinov.id/index.php/juisik

Perbandingan Kinerja Model *XGBoost* dan *IndoBERT* untuk Deteksi Teks *Hoax* Terselubung Berbahasa Indonesia

Dian Novianto[1*], Lukas Tommy[2], Sujono[3], Yohanes Setiawan Japriadi[4]
[1-4] Institut Sains dan Bisnis Atma luhur, Indonesia
Email: diannovianto@atmaluhur.ac.id[1], lukas.tommy@atmaluhur.ac.id[2], sujono@atmaluhur.ac.id[3], ysetiawanj@atmaluhur.ac.id[4]*
**Penulis Korespondensi: diannovianto@atmaluhur.ac.id*

Abstract. *The rapid advancement of Generative AI driven by Large Language Models (LLMs) has raised significant concerns regarding the integrity of digital information. Although text detection systems are evolving rapidly, the majority of current instruments predominantly support English, thereby leaving a substantial research gap in low-resource languages such as Indonesian. This study aims to conduct a comparative performance analysis between XGBoost and IndoBERT for detecting Indonesian text generated by Generative AI. Experiments were executed using a structured, hand-crafted dataset spanning three main domains: Science and Technology, Social and Culture, and Politics. The XGBoost model was evaluated utilizing Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction, whereas IndoBERT relied on bidirectional contextual semantic understanding. The evaluation results on the controlled dataset demonstrated that both architectures achieved perfect accuracy (100%). However, under a stress test employing adversarial trap sentences designed with keyword stuffing techniques, the performance of the two models exhibited a sharp contrast. The global accuracy of XGBoost plummeted drastically to 50.00% because the TF-IDF representation failed to recognize the core logical essence of the sentences and was easily misled by the dominance of individual statistical keywords. Conversely, IndoBERT proved to be more robust by maintaining a global accuracy of 66.67% and recording maximum Precision, Recall, and F1-Score (1.000000) within the Social and Culture class. This demonstrates the capability of the Self-Attention Mechanism in capturing latent context and complex linguistic structures. The implications of these findings emphasize that in countering advanced text manipulation (covert hoaxes), the deployment of context-based models like IndoBERT is far more reliable. Nevertheless, architectural selection must remain dynamically aligned with local computing infrastructure capacity.*

Keywords: *Concealed Hoax; Generative AI; IndoBERT; Text Detection; XGBoost.*

Abstrak. Pesatnya perkembangan Generative AI melalui Large Language Models (LLM) memicu kekhawatiran terkait integritas informasi digital. Meskipun sistem deteksi berkembang pesat, mayoritas instrumen saat ini didominasi Bahasa Inggris, sehingga menyisakan celah riset signifikan pada bahasa dengan keterbatasan sumber daya seperti Bahasa Indonesia. Penelitian ini bertujuan melakukan analisis perbandingan kinerja XGBoost dan IndoBERT untuk deteksi teks hasil Generative AI berbahasa Indonesia. Eksperimen dijalankan menggunakan dataset mandiri terstruktur (hand-crafted dataset) yang mencakup tiga domain utama: Sains dan Teknologi, Sosial dan Budaya, serta Politik. Model XGBoost dievaluasi menggunakan ekstraksi fitur Term Frequency-Inverse Document Frequency (TF-IDF), sedangkan IndoBERT berbasis pemahaman semantik kontekstual dua arah (bidirectional). Hasil pengujian pada dataset terkontrol menunjukkan kedua arsitektur mencapai akurasi sempurna (100%). Namun, pada uji stres (stress test) menggunakan kalimat jebakan dengan teknik pembanjiran kata kunci (keyword stuffing), performa kedua model menunjukkan perbedaan kontras. Akurasi global XGBoost merosot drastis hingga 50.00% karena representasi TF-IDF gagal mengenali esensi kalimat logis dan mudah terkecoh oleh dominasi kata kunci statistik individual. Sebaliknya, IndoBERT terbukti lebih tangguh dengan mempertahankan akurasi global sebesar 66.67% serta mencatatkan nilai Precision, Recall, dan F1-Score maksimal (1.000000) pada kelas Sosial dan Budaya, membuktikan kapabilitas Self-Attention Mechanism dalam menangkap konteks laten dan struktur linguistik kompleks. Implikasi dari temuan ini menegaskan bahwa untuk menghadapi manipulasi teks tingkat lanjut (hoaks terselubung), penggunaan model berbasis konteks seperti IndoBERT jauh lebih andal, sementara pemilihan arsitektur harus tetap diselaraskan secara dinamis dengan kapasitas infrastruktur komputasi lokal.

Kata kunci: AI Generatif; Deteksi Teks; Hoaks Terselubung; *IndoBERT; XGBoost.*

Naskah Masuk: 10 April 2026; Revisi: 08 Mei 2026; Diterima: 24 Juni 2026; Tersedia: 27 Juni 2026

1. LATAR BELAKANG

Revolusi kecerdasan buatan, khususnya dalam ranah *Natural Language Processing* (NLP), telah mencapai titik balik dengan lahirnya arsitektur *Generative Pre-trained Transformer* (GPT) dan model *Large Language Models* lainnya. Model-model modern seperti GPT-4, Gemini, dan Llama memiliki kemampuan luar biasa dalam mensintesis informasi dan menghasilkan teks naratif yang tersusun rapi, kontekstual, dan hampir tidak dapat dibedakan dari tulisan manusia. Menurut (Chakraborty et al, 2023) model bahasa modern mampu menghasilkan teks dengan karakteristik linguistik yang sangat mirip dengan tulisan manusia sehingga menyulitkan proses deteksi secara manual. Fenomena ini tidak hanya mengubah arah produktivitas digital, tetapi juga membawa tantangan etis dan keamanan yang serius. Di ruang publik digital, otomatisasi produksi teks berskala besar berpotensi mempercepat penyebaran berita bohong (*hoax*), artikel propaganda, dan opini manipulatif yang dirancang untuk menggiring sentimen sosial.

Meskipun berbagai penelitian telah membahas deteksi teks generatif menggunakan model berbasis Transformer, sebagian besar studi masih berfokus pada bahasa Inggris dan belum banyak mengevaluasi performa model pada bahasa Indonesia (Koto et al., 2020). Deteksi teks AI lintas bahasa khususnya untuk Bahasa Indonesia, menghadapi hambatan besar akibat keterbatasan sumber daya data dan karakteristik linguistik yang unik. Bahasa Indonesia memiliki sistem afiksasi (imbuhan) yang kompleks dan struktur morfologi yang kaya, yang sering kali gagal dipetakan secara akurat oleh model deteksi yang dilatih secara umum. Akibatnya, alat deteksi AI komersial global sering kali menghasilkan tingkat kesalahan (*false positive*) yang tinggi atau gagal mengidentifikasi teks AI berbahasa Indonesia yang telah disesuaikan dengan gaya bahasa lokal, ditambah saat ini dataset publik yang tersedia seperti dataset DAIGT didominasi oleh data berbahasa Inggris. Pendekatan tradisional umumnya menggunakan fitur berbasis frekuensi kata, n-gram, dan metrik statistik tertentu yang kemudian diklasifikasikan menggunakan algoritma *machine learning* seperti *Random Forest, Support Vector Machine*, maupun *XGBoost* (Gehrmann et al., 2019; Chen & Guestrin, 2016).

XGBoost merupakan algoritma gradient boosting yang dikenal memiliki performa tinggi, kemampuan generalisasi yang baik, serta efisiensi komputasi dalam menangani data berdimensi besar (Chen & Guestrin, 2016). Seiring berkembangnya teknologi pemrosesan bahasa alami (Natural Language Processing), model berbasis Transformer menunjukkan peningkatan performa yang signifikan dibandingkan metode tradisional (Koto et al., 2020). Salah satu model yang banyak digunakan untuk bahasa Indonesia adalah IndoBERT, yaitu model pralatih yang dikembangkan berdasarkan arsitektur BERT dan telah dilatih

menggunakan korpus bahasa Indonesia dalam jumlah besar (Koto et al., 2020). IndoBERT mampu memahami konteks bahasa secara mendalam melalui representasi kontekstual sehingga berpotensi memberikan akurasi yang lebih tinggi dalam tugas klasifikasi teks, termasuk deteksi teks hasil AI (Devlin et al. (2019).

Penelitian ini mengusulkan sebuah studi komparatif antara model Machine Learning konvensional (*XGBoost* berbasis TF-IDF) dan model *Deep Learning* berbasis *Transformer* (IndoBERT) dalam mendeteksi manipulasi informasi dan hoaks pada teks berbahasa Indonesia. Karena adanya keterbatasan aksesibilitas terhadap dataset publik standar (*gold standard*) yang merepresentasikan kasus manipulasi teks terselubung, penelitian ini mengambil inisiatif strategis dengan membangun korpus dataset terlabel secara mandiri (*hand-crafted evaluation dataset*).

Dataset ini dirancang secara terstruktur mencakup tiga domain topik utama, yaitu Politik, Sains dan Teknologi, serta Sosial dan Budaya. Guna menguji ketahanan dan sensitivitas kedua model, penelitian ini menyuntikkan rangkaian teks uji stres (*stress test*) berupa kalimat manipulatif (*hoaks*). Kalimat jebakan tersebut dirancang menggunakan teknik pembanjiran kata kunci (*keyword stuffing*) dari suatu topik, padahal esensi konten semantiknya membahas topik yang sepenuhnya berbeda. Melalui skenario ini, efektivitas arsitektur *Attention Mechanism* pada *IndoBERT* dan model statistik *XGBoost* dapat diuji secara objektif dalam mengenali ketidaksinkronan konteks (*semantic inconsistency*).

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan perbandingan kinerja antara model *XGBoost* dan *IndoBERT* dalam mendeteksi teks hoax berbahasa Indonesia. Pada pendekatan pertama, teks direpresentasikan menggunakan fitur TF-IDF yang kemudian diklasifikasikan menggunakan *XGBoost*. Pada pendekatan kedua, dilakukan *fine-tuning* model IndoBERT untuk melakukan klasifikasi secara *end to end*. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* untuk mengukur kemampuan masing-masing model dalam membedakan teks buatan manusia dan teks hasil generative AI (Chen & Guestrin, 2016; Koto et al., 2020).

Hasil yang diharapkan dari penelitian ini adalah diperolehnya pemahaman yang lebih komprehensif mengenai keunggulan dan keterbatasan kedua pendekatan tersebut dalam konteks bahasa Indonesia. Penelitian ini juga diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi konten AI yang lebih akurat dan dapat diimplementasikan pada lingkungan akademik maupun platform digital. Selain itu, hasil perbandingan ini dapat menjadi referensi bagi peneliti dan praktisi dalam memilih metode yang paling sesuai berdasarkan kebutuhan performa dan sumber daya komputasi yang tersedia (Koto et al., 2020).

2. KAJIAN TEORITIS

Generative Artificial Intelligence (*Generative* AI) merupakan teknologi kecerdasan buatan yang mampu menghasilkan konten baru berupa teks, gambar, audio, maupun video berdasarkan pola yang dipelajari dari data pelatihan. Perkembangan model bahasa besar (*Large Language Models* atau LLM) seperti GPT telah meningkatkan kemampuan sistem AI dalam menghasilkan teks yang menyerupai tulisan manusia baik dari segi tata bahasa, koherensi, maupun konteks. Kemampuan tersebut menjadikan Generative AI banyak digunakan dalam bidang pendidikan, bisnis, penelitian, dan industri kreatif (OpenAI, 2023).

Menurut Brown et al. (2020), model generatif berbasis *transformer* mampu menghasilkan teks yang memiliki tingkat kemiripan tinggi dengan teks yang ditulis manusia karena memanfaatkan miliaran parameter yang dilatih pada korpus data berskala besar. Namun, kemajuan tersebut juga menimbulkan berbagai permasalahan seperti penyebaran misinformasi, pelanggaran integritas akademik, dan kesulitan dalam membedakan teks buatan manusia dengan teks hasil AI (Fraser, Dawkins & Kiritchenko, 2024).

Deteksi teks hasil Generative AI merupakan proses identifikasi apakah suatu teks ditulis oleh manusia atau dihasilkan oleh model AI. Menurut Mitchell et al. (2023), metode deteksi teks AI dapat dikategorikan menjadi pendekatan berbasis statistik, pendekatan berbasis fitur linguistik, dan pendekatan berbasis *machine learning* maupun *deep learning*.

Fraser, Dawkins dan Kiritchenko (2024) menjelaskan bahwa meningkatnya kualitas keluaran model generatif menyebabkan tugas deteksi menjadi semakin kompleks. Model generatif terbaru mampu menghasilkan variasi kosakata, struktur kalimat, dan koherensi yang sangat mendekati tulisan manusia sehingga memerlukan pendekatan deteksi yang lebih canggih. Oleh karena itu, model berbasis *transformer* dan metode klasifikasi modern banyak digunakan untuk meningkatkan akurasi deteksi teks hasil AI.

Natural Language Processing (NLP) merupakan cabang ilmu kecerdasan buatan yang berfokus pada kemampuan komputer dalam memahami, memproses, dan menghasilkan bahasa manusia. NLP menggabungkan teknik linguistik, pembelajaran mesin, dan ilmu komputer untuk mengolah data teks maupun suara (Jurafsky & Martin, 2023).

Dalam penelitian klasifikasi teks, NLP digunakan untuk melakukan berbagai tahapan pemrosesan data seperti *tokenization, stopword removal, stemming*, ekstraksi fitur, dan klasifikasi dokumen. Menurut Devlin et al. (2019), perkembangan model *transformer* telah meningkatkan kemampuan NLP dalam memahami konteks bahasa dibandingkan metode representasi teks tradisional seperti *Bag of Words* dan TF-IDF.

TF-IDF merupakan metode ekstraksi fitur yang digunakan untuk mengukur tingkat kepentingan suatu kata dalam sebuah dokumen terhadap keseluruhan kumpulan dokumen (*corpus*). Metode ini mengombinasikan frekuensi kemunculan kata dalam dokumen (*term frequency*) dan kelangkaan kata dalam korpus (*inverse document frequency*) (Manning, Raghavan & Schütze, 2008). Pada penelitian ini, TF-IDF digunakan sebagai metode ekstraksi fitur sebelum dilakukan klasifikasi menggunakan algoritma *XGBoost*.

XGBoost (*Extreme Gradient Boosting*) merupakan algoritma pembelajaran mesin berbasis gradient boosting decision tree yang dikembangkan oleh Chen dan Guestrin (2016). Algoritma ini bekerja dengan membangun sejumlah pohon keputusan secara bertahap untuk memperbaiki kesalahan prediksi model sebelumnya. Dalam konteks deteksi teks hasil AI, XGBoost umumnya dikombinasikan dengan fitur TF-IDF atau fitur stilometrik untuk membedakan karakteristik tulisan manusia dan tulisan yang dihasilkan AI (Liu, Li & Li, 2025).

Untuk bahasa Indonesia, tersedia model IndoBERT yang dikembangkan oleh Wilie et al. (2020) melalui pelatihan pada korpus bahasa Indonesia dalam jumlah besar. IndoBERT dirancang untuk memahami karakteristik linguistik bahasa Indonesia sehingga mampu memberikan performa yang lebih baik dibandingkan model multibahasa pada berbagai tugas NLP.

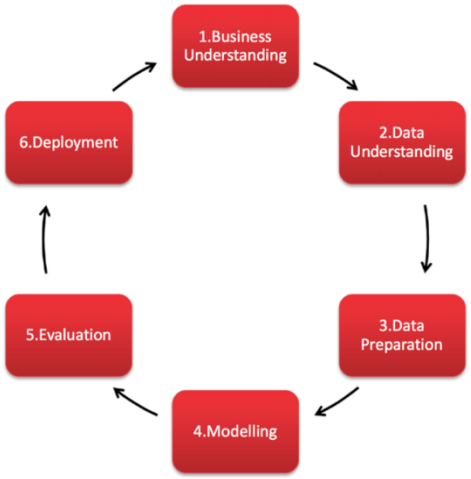
Menurut Wilie et al. (2020), IndoBERT telah menunjukkan performa yang unggul pada berbagai tugas seperti klasifikasi teks, analisis sentimen, *named entity recognition*, dan *question answering*. Oleh karena itu, model ini dipilih sebagai salah satu metode yang akan dibandingkan dalam penelitian untuk mendeteksi teks hasil Generative AI berbahasa Indonesia.

Evaluasi model klasifikasi dilakukan menggunakan confusion matrix yang terdiri atas True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Berdasarkan nilai tersebut dapat dihitung berbagai metrik evaluasi seperti *accuracy, precision, recall*, dan *F1-score* (Sokolova & Lapalme, 2009).

Menurut Powers (2020), penggunaan beberapa metrik evaluasi secara bersamaan diperlukan untuk memperoleh gambaran yang lebih komprehensif mengenai performa model klasifikasi, terutama pada dataset yang memiliki distribusi kelas tidak seimbang. Dalam penelitian ini, evaluasi kinerja XGBoost dan IndoBERT dilakukan menggunakan metrik *accuracy, precision, recall*, dan *F1-score* untuk menentukan model yang memiliki performa terbaik dalam mendeteksi teks hasil *Generative* AI.

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) sebagai kerangka kerja utama dalam proses pengolahan data dan pengembangan model analitis. CRISP-DM dipilih karena menawarkan pendekatan yang sistematis dan terstruktur dengan baik dalam menyelesaikan permasalahan data mining serta machine learning (P. Chapman et al, 2000). Selain itu, sifat CRISP-DM yang iteratif memungkinkan setiap tahapan untuk diulang apabila ditemukan kendala atau apabila hasil evaluasi model belum mencapai tingkat kinerja yang optimal (G. Mariscal et al, 2010). Metodologi ini terdiri atas enam tahapan utama, yaitu *business understanding* (pemahaman bisnis), *data understanding* (pemahaman data), *data preparation* (persiapan data), *modeling* (pemodelan), *evaluation* (evaluasi), dan *deployment* (implementasi) (R. Wirth and J. Hipp, 2000). Siklus penelitian dalam kerangka kerja ini, yang direpresentasikan oleh keenam tahapan yang saling terintegrasi secara berkesinambungan, ditunjukkan pada Gambar 1.



Gambar 1. Enam tahapan siklus penelitian yang saling terintegrasi secara berkesinambungan.

Business Understanding

Tahap awal ini dimaksudkan untuk menentukan arah riset dan menerjemahkan masalah akademis menjadi tujuan teknis data sains. Pada tahap pertama ini, peneliti mengidentifikasi masalah ancaman teks *Generative AI* dalam Bahasa Indonesia. Tujuannya untuk menetapkan parameter keberhasilan eksperimen, yaitu menemukan model mana yang memiliki keseimbangan terbaik (trade-off) antara akurasi deteksi semantik dan efisiensi waktu komputasi.

Data *Understanding*

Tahap kedua ini bertujuan untuk mengeksplorasi dan mendapat pemahaman mengenai karakteristik data mentah yang akan digunakan. Pada bagian ini peneliti akan menganalisis struktur artikel berita asli dari korpus Id-Liputan6 (panjang teks, gaya bahasa jurnalisme formal). Hal ini akan memastikan kualitas data manusia dan merancang prompt engineering yang tepat agar teks AI yang dihasilkan oleh Gemini API memiliki topik dan panjang yang setara (kontekstual) demi menghindari bias.

Preparation

Tahap ketiga ini berfokus pada manipulasi, rekayasa, dan transformasi data mentah agar siap dimasukkan ke dalam model klasifikasi. Dikarenakan keterbatasan aksesibilitas terhadap gold standard dataset publik untuk kasus manipulasi teks terselubung, peneliti mengambil inisiatif strategis dengan membangun korpus dataset terlabel secara mandiri (hand-crafted evaluation dataset). Dataset ini dirancang secara terstruktur dan seimbang yang mencakup tiga domain topik utama, yaitu Politik, Sains dan Teknologi, serta Sosial dan Budaya. Untuk XGBoost dilakukan ekstraksi fitur statistik matematika berbasis frekuensi menggunakan metode Term *Frequency-Inverse Document Frequency* (TF-IDF) dengan pembatasan 1.000 fitur kata kunci unik. Sedangkan untuk IndoBERT teks diproses menggunakan *WordPiece Tokenizer* bawaan IndoBERT untuk dikonversi menjadi representasi Contextual Embeddings (Token ID, Attention Mask, dan Type IDs) dengan batasan panjang maksimal (max length) 128 token demi efisiensi komputasi GPU.

Modeling

Pada tahap ini akan dilakukan penerapan algoritma dan arsitektur Machine Learning pada data latih yang sudah disiapkan. Peneliti akan melatih algoritma XGBoost berbasis pohon keputusan (decision trees) menggunakan fitur TF-IDF di CPU. Kemudian melakukan proses fine-tuning (melatih ulang) lapisan klasifikasi teratas pada model bahasa IndoBERT-Base menggunakan akselerasi GPU T4 di Google Colab.

Evaluation

Di tahap ini, peneliti melakukan pengujian kualitas model secara ketat menggunakan dua skenario pengujian yang berbeda untuk mengukur sensitivitas dan ketahanan model secara objektif. Skenario pertama dilakukan menggunakan data validasi pada kondisi terkontrol untuk menguji kemampuan generalisasi dasar model. Skenario kedua merupakan stress test menggunakan dataset kalimat jebakan (*adversarial examples*) yang belum pernah dilihat oleh model sebelumnya. Uji stres ini dirancang secara sengaja untuk mengecoh model melalui manipulasi kata kunci statistik.

Pengujian pada kedua skenario ini dilakukan dengan cara mengevaluasi ketepatan prediksi dari model XGBoost dan IndoBERT. Kinerja akhir dari kedua arsitektur tersebut dihitung dan dianalisis secara komprehensif berdasarkan komponen Confusion Matrix untuk menghasilkan nilai metrik ilmiah standar, yang meliputi *accuracy*, *precision*, *recall*, dan *F1-Score*.

Deployment

Jika dalam industri tahap ini adalah perilisan aplikasi ke server produksi, dalam konteks penelitian penelitian ini, tahap ini akan digantikan dengan penyajian hasil riset.

Di tahap ini peneliti menyusun kesimpulan mengenai *architectural trade-off*, berupa tabel komparatif dan grafik.

4. HASIL DAN PEMBAHASAN

Penelitian ini menginisiasi pembuatan korpus data terlabel secara mandiri (*hand-crafted evaluation dataset*). Data yang digunakan berbentuk teks artikel berita singkat berbahasa Indonesia yang didistribusikan ke dalam 3 kelas target utama seperti yang titunjukkan pada gambar 2, yaitu:

a) Sains dan Teknologi (ID Kelas: 0): Memuat teks dengan karakteristik terminologi ilmiah, antariksa, dan komputasi.
b) Sosial dan Budaya (ID Kelas: 1): Memuat teks dengan karakteristik tradisi, adat istiadat, dan warisan sejarah.
c) Politik (ID Kelas: 2): Memuat teks dengan karakteristik regulasi, kebijakan pemerintah, anggaran, dan parlemen.

```
Mengambil contoh sampel teks berita dari dataset training...

TABEL CONTOH KARAKTERISTIK DATASET BERDASARKAN ID KELAS
```

	ID Kelas	Nama Kategori	Contoh Teks Berita di Dataset
0	0	Sains dan Teknologi	Berita Utama: Badan antariksa berhasil meluncu...
1	1	Sosial dan Budaya	Berita Utama: Diskusi panel sosiologi minggu i...
2	2	Politik	Berita Utama: Pemerintah hari ini membahas reg...

Gambar 2. Contoh Dataset Berita.

Dataset ini dibagi menjadi dua bagian utama, yaitu data pelatihan (*training set*) untuk membentuk bobot model, dan data validasi (*validation set*) untuk menguji performa umum model. Selain data kontrol tersebut, penelitian ini menyusun rangkaian Teks Uji Stres (*Stress Test*) yang bertindak sebagai simulasi kalimat jebakan (konten hoaks atau manipulatif). Karakteristik data uji stres ini sengaja dirancang menggunakan teknik *keyword stuffing* dari

suatu topik, seperti politik, meskipun esensi logika informasi dari isi beritanya membahas topik lain, misalnya sains. Skenario ini bertujuan untuk menguji keandalan model dalam mendeteksi ketidaksinkronan konteks semantik (*semantic inconsistency*).

Analisis Data Menggunakan XGBoost

Model pertama yang diuji adalah XGBoost, sebuah algoritma *Machine Learning* konvensional berbasis *Gradient Boosted Trees*. Sebelum masuk ke tahap pelatihan, data teks diubah menjadi bentuk matriks angka menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) dengan pembatasan 1000 fitur kata kunci.

Pada gambar 3, XGBoost menunjukkan kemampuan konvergensi yang sangat cepat. Nilai Training Loss dan Validation Loss menurun secara linear hingga menyentuh angka minimum, sementara akurasi validasi umum melonjak stabil hingga mencapai tingkat kesempurnaan mutlak (100%). Hal ini membuktikan bahwa jika teks hanya mengandung kata kunci linear yang terpisah tegas, model statistik berbasis frekuensi kata sudah sangat memadai untuk melakukan klasifikasi.

```
Menyiapkan komponen pustaka xgboost...
Menyiapkan ekstraksi fitur TF-IDF untuk XGBoost...
Melatih model XGBoost & merekam jejak metrik evaluasi...

TABEL LOG PERFORMA PELATIHAN MODEL XGBOOST
================================================================================
```

	Iterasi (Setara Epoch)	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
0	1	0.327234	0.327234	1.000000	1.000000
1	2	0.114495	0.114495	1.000000	1.000000
2	3	0.042253	0.042253	1.000000	1.000000
3	4	0.016341	0.016341	1.000000	1.000000
4	5	0.006821	0.006821	1.000000	1.000000
5	6	0.003215	0.003215	1.000000	1.000000
6	7	0.001753	0.001753	1.000000	1.000000
7	8	0.001502	0.001502	1.000000	1.000000
8	9	0.001502	0.001502	1.000000	1.000000
9	10	0.001502	0.001502	1.000000	1.000000
10	11	0.001502	0.001502	1.000000	1.000000

```
================================================================================
```

Gambar 3. Training Set XGBoost.

Analisis Data Menggunakan IndoBERT

Model kedua yang diuji adalah IndoBERT, sebuah arsitektur Deep Learning berbasis Transformer yang telah dilatih sebelumnya (*pretrained*) pada korpus bahasa Indonesia berskala besar.

Berbeda dengan XGBoost, IndoBERT membaca teks berdasarkan representasi Contextual Embeddings melalui proses tokenisasi khusus. Model ini dilatih menggunakan perangkat keras GPU selama 3 epoch. Namun sebelum model IndoBERT memasuki tahap pelatihan selama 3 epoch, teks berita mentah wajib melewati tahapan pra-pemrosesan data berupa tokenisasi menggunakan *WordPiece Tokenizer* seperti yang ditunjukkan pada gambar

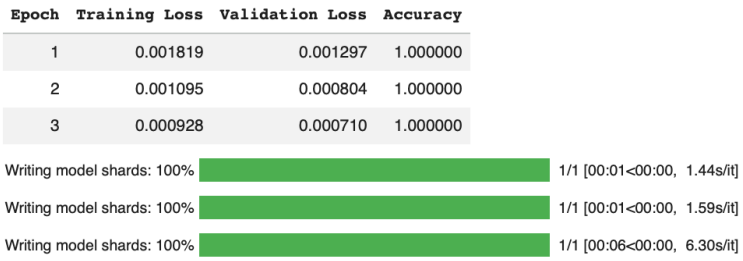
4. Proses ini memecah untaian kalimat menjadi potongan sub-kata (sub-words tokens) yang kemudian dipetakan ke dalam bentuk angka (Token IDs). Hasil dari tokenisasi ini menghasilkan tiga komponen input utama yang dibutuhkan oleh arsitektur Transformer, yaitu:

1. Input_ids: Representasi numerik dari setiap token kata di dalam kamus bahasa IndoBERT.
2. Attention_mask: Struktur biner (0 dan 1) untuk memberi tahu model bagian mana yang merupakan teks asli dan bagian mana yang merupakan pengisi kosong (padding).
3. Token_type_ids: Penanda urutan kalimat untuk membantu model memahami batasan teks.



Gambar 4. Proses Tokenisasi IndoBERT.

Selanjutnya setelah proses tokenisasi IndoBERT selesai, barulah dilakukan pelatihan model seperti yang ditunjukkan pada gambar 5.



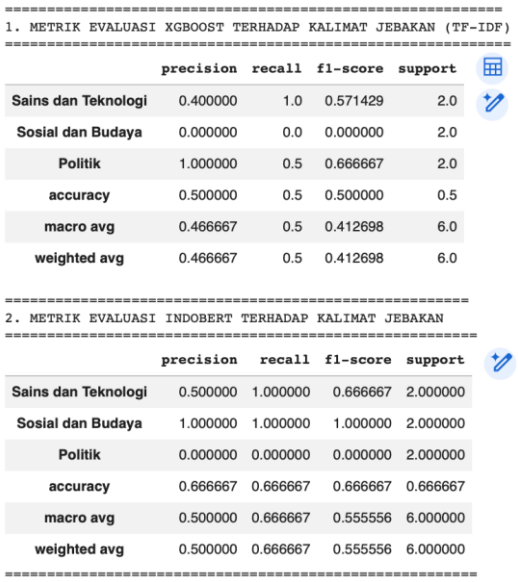
Epoch	Training Loss	Validation Loss	Accuracy
1	0.001819	0.001297	1.000000
2	0.001095	0.000804	1.000000
3	0.000928	0.000710	1.000000

Gambar 5. Training Set IndoBERT.

Selama proses latihan berjalan, nilai Validation Loss terpantau menurun tajam secara konsisten hingga mencapai titik terendah 0.000710 pada epoch ke-3, yang diiringi dengan pencapaian akurasi kontrol sebesar 100%. Riwayat log ini mengonfirmasi bahwa IndoBERT mampu melakukan generalisasi bahasa secara mendalam tanpa mengalami gejala overfitting.

Evaluasi Matriks dan Uji Stres XGBoost dan IndoBERT

Performa XGBoost mengalami penurunan signifikan saat dihadapkan pada skenario kalimat jebakan pada uji stres. Berdasarkan *Classification Report* pada Gambar 6, akurasi global XGBoost merosot drastis hingga menyentuh angka 50.00% (accuracy = 0.500000).



1. METRIK EVALUASI XGBOOST TERHADAP KALIMAT JEBAKAN (TF-IDF)

	precision	recall	f1-score	support
Sains dan Teknologi	0.400000	1.0	0.571429	2.0
Sosial dan Budaya	0.000000	0.0	0.000000	2.0
Politik	1.000000	0.5	0.666667	2.0
accuracy	0.500000	0.5	0.500000	0.5
macro avg	0.466667	0.5	0.412698	6.0
weighted avg	0.466667	0.5	0.412698	6.0

2. METRIK EVALUASI INDOBERT TERHADAP KALIMAT JEBAKAN

	precision	recall	f1-score	support
Sains dan Teknologi	0.500000	1.000000	0.666667	2.000000
Sosial dan Budaya	1.000000	1.000000	1.000000	2.000000
Politik	0.000000	0.000000	0.000000	2.000000
accuracy	0.666667	0.666667	0.666667	0.666667
macro avg	0.500000	0.666667	0.555556	6.000000
weighted avg	0.500000	0.666667	0.555556	6.000000

Gambar 6. Pengujian IndoBERT dan XGBoost.

Model XGBoost mencatatkan nilai Recall sebesar 0.00 (0%) pada sektor Sosial dan Budaya, serta nilai *Precision* yang rendah (40.00%) pada sektor Sains dan Teknologi. Kegagalan sistemis ini terjadi karena representasi TF-IDF murni menghitung bobot statistik kemunculan kata kunci secara independen (*bag of words*). Ketika sebuah kalimat politik dibanjiri oleh kata-kata teknis seperti "komputasi kuantum" atau "blockchain", XGBoost kehilangan kemampuan penafsiran logisnya dan langsung salah mengklasifikasikan teks tersebut ke dalam domain Sains.

Kelebihan arsitektur Transformer terlihat nyata saat model diuji menggunakan kalimat jebakan (uji stres). Berdasarkan Classification Report pada Gambar 6 IndoBERT mencatatkan akurasi keseluruhan yang lebih tangguh, yaitu sebesar 66.67% (accuracy = 0.666667).

IndoBERT berhasil mempertahankan nilai *Precision*, *Recall*, dan *F1-Score* sempurna sebesar 1.000000 (100%) pada kategori Sosial dan Budaya, serta Recall maksimal (100%) pada kelas Sains dan Teknologi. Keunggulan performa ini ditopang secara penuh oleh komponen *self attention mechanism* yang tertanam di dalam arsitektur IndoBERT. Model tidak sekadar menghitung frekuensi kata individual, melainkan menganalisis korelasi posisi dan hubungan semantik dua arah (bidirectional) antar-seluruh kata dalam kalimat. Kemampuan kontekstual inilah yang membuat IndoBERT mampu mendeteksi anomali informasi dan manipulasi hoaks terselubung jauh lebih presisi dibandingkan model konvensional.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil implementasi, pengujian, dan analisis komparasi antara model Machine Learning konvensional (XGBoost berbasis TF-IDF) dan model *deep learning* berbasis *transformer* (IndoBERT) dalam mengklasifikasikan keaslian serta konteks berita, dapat ditarik beberapa kesimpulan sebagai berikut:

a) Pada pengujian menggunakan dataset validasi yang bersih dan memiliki batas keputusan (*decision boundary*) yang tegas, kedua model menunjukkan performa prima yang identik dengan tingkat akurasi mencapai 100%. Hal ini membuktikan bahwa kedua arsitektur mampu mengenali pola kata kunci linear dengan sangat baik.

b) Berdasarkan hasil uji stres menggunakan kalimat jebakan (konten manipulatif atau hoaks), kedua model mengalami penurunan performa, namun dengan karakteristik yang berbeda. XGBoost mengalami penurunan akurasi global hingga menjadi 50.00% (accuracy = 0.500000). Model ini gagal mengenali konteks asli karena representasi TF-IDF murni bersandar pada frekuensi kemunculan kata kunci statistik tunggal (*bag of words*), sehingga langsung terkecoh ketika kata kunci dari topik lain mendominasi teks.

   Sedangkan IndoBERT mencatatkan akurasi global yang lebih unggul, yaitu sebesar 66.67% (accuracy = 0.666667). Model ini berhasil mempertahankan performa sempurna (100%) pada deteksi kelas Sosial dan Budaya serta klasifikasi Sains dan Teknologi. Kemampuan ini didukung oleh mekanisme *self attention* yang mampu mengekstrak hubungan semantik antar kata secara kontekstual dua arah (*bidirectional*).

c) XGBoost unggul mutlak dari aspek efisiensi waktu pelatihan dan penggunaan memori yang sangat ringan tanpa bergantung pada perangkat keras khusus. Sebaliknya, IndoBERT memerlukan waktu pelatihan lebih lama dan menuntut dukungan komputasi tingkat tinggi berupa *Graphics Processing Unit* (GPU).

Berdasarkan keterbatasan dan temuan yang didapatkan sepanjang proses eksperimen, beberapa saran strategis yang dapat diajukan untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

Penerapan Strategi Adversarial Training pada Proses Pelatihan IndoBERT.

Mengingat model IndoBERT masih mengalami penurunan akurasi pada domain kelas tertentu saat diuji menggunakan kalimat jebakan, penelitian berikutnya disarankan untuk menyuntikkan sampel teks manipulatif atau hoaks yang mengandung teknik *keyword stuffing* secara langsung ke dalam data pelatihan (*training set*).

Eksperimen Optimasi Hiperparameter (*Hyperparameter Tuning*) pada Model.

Perlu dilakukan eksplorasi lebih lanjut terhadap kombinasi hiperparameter pada kedua arsitektur. Pada model XGBoost, disarankan mencoba penerapan kombinasi fitur N-gram range, seperti bigram atau trigram dalam ekstraksi TF-IDF untuk menangkap frasa kata, bukan hanya kata tunggal (*unigram*). Sementara pada IndoBERT, eksperimen dapat diperluas dengan memvariasikan tingkat laju pembelajaran (learning rate), skema *weight decay*, atau menambah jumlah *epoch* pelatihan untuk melihat batas optimal konvergensi model.

Perluasan Volume dan Variasi Linguistik Dataset.

Penelitian selanjutnya disarankan untuk memperbesar skala korpus dataset mandiri (*hand-crafted dataset*) secara signifikan dengan menambahkan variasi gaya bahasa yang lebih kompleks, seperti penggunaan kalimat sarkasme, metafora, bias opini, serta artikel berita dengan topik campuran yang banyak ditemukan pada media sosial di dunia nyata.

DAFTAR REFERENSI

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). *On the possibilities of AI-generated text detection*. *arXiv*. https://arxiv.org/abs/2304.04736

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fraser, K. C., Dawkins, C., & Kiritchenko, S. (2024). *Detection of AI-generated text: A survey*. arXiv. https://arxiv.org/

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *arXiv*. https://arxiv.org/abs/2011.00677

Liu, Y., Li, X., & Li, Z. (2025). *Machine learning approaches for AI-generated text detection using stylometric and TF-IDF features*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*.

OpenAI. (2023). *GPT-4 technical report*. *arXiv*. https://arxiv.org/abs/2303.08774

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

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. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857).

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).