

BAB 2

LANDASAN TEORI

2.1 Keamanan Siber (Cyber Security)

2.1.1 Definisi Keamanan Siber

Dalam era teknologi sekarang, banyak sekali data sensitif / berharga yang disimpan dalam bentuk digital. Data ini dapat berupa data pribadi, aset digital, data sensitif suatu perusahaan. Dengan semakin majunya perkembangan jaman, maka semakin banyak orang yang beralih dalam menyimpan data dalam dunia digital, kondisi ini memicu peningkatan target oleh pelaku kejahatan siber terhadap data-data digital tersebut. Dalam upaya untuk melindungi data - data berharga tersebut maka diciptakanlah keamanan siber untuk melindungi data tersebut. Keamanan siber (Cybersecurity) merupakan praktik melindungi aset digital berupa komputer, jaringan, aplikasi, sistem, dan data dari serangan [12]. Pada dasarnya, keamanan siber melibatkan serangkaian tindakan berupa pencegahan, pendeteksian, dan respons terhadap sebuah serangan. praktik keamanan siber tidak hanya sebatas ditingkat individu, namun juga diimplementasikan oleh perusahaan. Keamanan siber memastikan perlindungan data-data sensitif dari akses yang tidak sah.

2.1.2 Ancaman dan Serangan dalam Keamanan Siber

Perkembangan pesat era digital tidak hanya menghadirkan kemudahan, namun juga membuka celah bagi berbagai taktik serangan untuk mengakses data sensitif. Dalam ranah keamanan siber, jenis serangan sangat beragam, mulai dari upaya akses tidak sah hingga malware yang merusak dan menghancurkan data. Selama internet eksis, bentuk-bentuk serangan ini akan terus berevolusi, “berlomba” dengan mekanisme pertahanan dalam menutup setiap celah kelemahan [13, 14].

Adapun berbagai jenis dari bentuk ancaman dan serangan siber, yaitu:

1. **Malware (Malicious Software):** Malware adalah perangkat lunak berbahaya, mencakup virus, ransomware, trojan, dan spyware. Tujuannya adalah mendapatkan akses tidak sah atau merusak data—seringkali dengan enkripsi paksa (ransomware) atau pencurian informasi (spyware) [13].

2. Denial of Service (DoS / DDoS): Serangan DoS atau DDoS membanjiri server target dengan permintaan palsu sehingga layanan menjadi tidak tersedia bagi pengguna sah. Teknik ini dapat melumpuhkan situs web, aplikasi, atau infrastruktur jaringan [13].
3. Phishing: Phishing memanfaatkan rekayasa sosial untuk menipu korban agar mengungkapkan informasi sensitif (kata sandi, nomor kartu)—biasanya melalui email, pesan singkat, atau situs palsu yang menyerupai aslinya [14].

Dengan berkelanjutannya perkembangan AI, maka akan terdapat ancaman dimana serangan siber kedepannya dapat memiliki fitur - fitur ai. Hal ini membuat diperlukannya pendekatan yang proaktif terhadap keamanan siber, yang melibatkan pemantauan berkelanjutan, pengumpulan informasi ancaman, dan pengembangan sistem pertahanan yang lebih canggih dalam menahan serangan.

2.1.3 Peran AI dan Machine Learning dalam Keamanan Siber

Pada awalnya, sistem keamanan siber pertama yang dikembangkan adalah *Initial Threat Detection: the Rule-Based System* pada tahun 1970. Sistem ini efektif dalam mendeteksi ancaman yang sudah diketahui sebelumnya, namun memiliki kelemahan signifikan dalam menghadapi serangan atau ancaman baru yang belum terdefinisi [15]. Perkembangan sistem keamanan siber terus berlanjut melalui beberapa tahap, termasuk kemampuan untuk mendeteksi anomali atau keanehan dalam sebuah sistem. Seiring dengan kemajuan teknologi yang pesat, era kecerdasan buatan (AI) dan Pembelajaran Mesin (ML) telah tiba, membawa perubahan signifikan dalam bidang keamanan siber.

Saat ini, AI dan ML telah menjadi alat yang esensial dalam ranah keamanan siber. Peran AI dalam keamanan siber terlihat dari kemampuannya dalam memproses data dan mengembangkan berbagai algoritma yang digunakan untuk pendeteksian ancaman, sehingga membantu pengembangan sistem keamanan yang lebih efektif [15]. Dengan kemampuannya memproses kumpulan dataset dalam jumlah yang sangat besar, AI/ML memungkinkan pembuatan sistem yang canggih dengan berbagai fitur, seperti: mendeteksi ancaman di tahap awal, melakukan analisis data secara *real-time*, memprediksi potensi serangan di masa depan, mengotomatisasi pencarian kerentanan atau potensi ancaman, dan mendeteksi anomali dalam perilaku sistem atau jaringan [15, 16].

Kemampuan-kemampuan ini memberikan beberapa keunggulan dalam

sistem pendeteksian ancaman, termasuk: tingkat akurasi yang lebih tinggi, skalabilitas yang lebih baik, waktu respons yang jauh lebih cepat terhadap insiden, dan pengurangan jumlah positif palsu (*false positive*) [16]. Kemajuan ini sangat membantu individu maupun perusahaan dalam memperkuat pertahanan mereka dan melindungi aset digital dari berbagai jenis ancaman siber. Melihat dinamika ancaman siber yang terus berkembang seiring dengan inovasi teknologi, dapat dipastikan bahwa peran dan pengembangan keamanan siber, khususnya yang berbasis AI/ML, akan terus menjadi prioritas di masa mendatang.

2.2 Konsep Dasar Deephoaks dan Dampaknya

Deephoaks merupakan penerapan teknologi *deepfake* yang digunakan untuk menyebarkan misinformasi dan menjadi ancaman besar dalam ranah keamanan siber [17]. *Deepfake* merupakan penerapan teknologi kecerdasan buatan (*deep learning*) yang digunakan untuk membuat konten sintetis dengan tingkat realisme tinggi, yang dapat berupa manipulasi teks, audio, gambar, dan video. Meskipun istilah *deepfake* sering kali diasosiasikan dengan manipulasi visual dan audio, teknologi generatif yang mendasarinya juga sangat relevan dalam penciptaan konten tekstual yang meyakinkan.

Dengan kecanggihan model generatif, dimungkinkan pembuatan konten yang menyerupai berita asli namun memiliki informasi palsu (hoaks) yang sulit dibedakan dari yang asli [18, 19]. Dalam konteks ini, *deepfake text* merujuk pada teks yang dihasilkan secara sintetis oleh model bahasa generatif seperti GPT-2/GPT-3, model ini telah dibuktikan dalam menghasilkan konten artikel palsu yang sangat meyakinkan [20, 21]. Menurut Jiameng et al., kemajuan model generatif memungkinkan mesin untuk menghasilkan teks palsu (termasuk *deepfake text*) dalam topik apa pun dengan kualitas yang tinggi [19]. Zellers et al. bahkan telah menunjukkan demonstrasi ancaman dari *deepfake text* dalam pembuatan artikel berita palsu yang meyakinkan dengan menggunakan model GPT-2 [22]. Gambar 2.1 menunjukkan peningkatan kasus penyebaran *deepfake* di Indonesia hingga sebesar 1550%.

VIDA catat penipuan “deepfake” di Indonesia melonjak 1.550 persen

Jumat, 1 November 2024 20:09 WIB waktu baca 2 menit



Co-founder dan Presiden VIDA Sati Rasuanto (kiri) menghadiri VIDA Executive Summit 2024. ANTARA/HO-VIDA.

Jakarta (ANTARA) - PT Indonesia Digital Identity (VIDA), penyelenggara Sertifikasi Elektronik (PSrE) yang terdaftar pada Kementerian Komunikasi dan Digital (Komdigi), mencatat terdapat lonjakan jumlah kasus penipuan menggunakan *deepfake* sebesar 1.550 persen di Indonesia pada 2022-2023.

Terpopuler

- Samsung gelar Galaxy Unpacked pada 9 Juli 2025, apa yang akan rilis? 22 jam lalu
- Aplikasi Whatsapp dilarang dari perangkat DPR AS 21 jam lalu
- Kemkomdigi sedang memblokir situs pemasar pulau kecil di Anambas 19 jam lalu
- Menkomdigi: ASN harus jadi penggerak transformasi digital 12 jam lalu
- Perplexity luncurkan aplikasi "browser" Comet untuk pengguna Windows 22 jam lalu

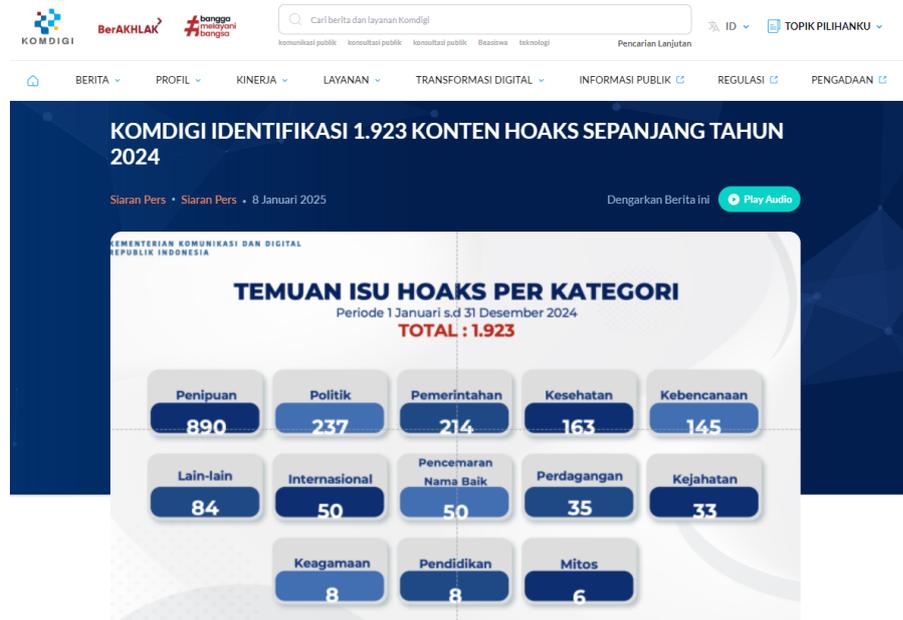
Top News



Gambar 2.1. Penipuan Deepfake Melonjak di Indonesia Sebesar 1550%

Kecanggihan model AI dalam menghasilkan teks telah mencapai titik di mana timbul tantangan serius dalam mengidentifikasi keaslian informasi daring. Pu dkk., (2022) menemukan bahwa tingkat akurasi manusia dalam membedakan teks asli dengan teks buatan GPT-3 sebesar 52% [21]. Kondisi ini menimbulkan kekawathiran publik terhadap kebenaran informasi yang disebar secara daring, dan menjadi celah bagi penyebaran berita hoaks yang dapat mengancam keamanan informasi dan siber.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA



Gambar 2.2. Isu penyebaran hoaks di Indonesia

Gambar 2.2 menunjukkan penyebaran isu hoaks yang terjadi di Indonesia dalam rangka waktu 1 tahun (2024). Isu hoaks yang muncul bervariasi dan mencakup ke berbagai topik dengan penipuan sebagai isu yang paling banyak terjadi [23].

2.3 Model BERT dan IndoBERT (Bidirectional Encoder Representations from Transformers)

2.3.1 Pengertian dan Prinsip Kerja BERT

BERT (Bidirectional Encoder Representations from Transformers) adalah model pembelajaran mesin yang berbasis transformer untuk representasi bahasa. Diperkenalkan oleh Devlin (2018), BERT dirancang dengan tujuan mempelajari representasi teks secara *bidirectional* (dua arah) sambil melihat konteks kiri dan kanan dari sebuah kata secara bersamaan di dalam semua lapisannya [11]. Dengan mengandalkan arsitektur transformer encoder bertingkat, BERT memiliki kemampuan untuk menangkap konteks global dari sebuah teks.

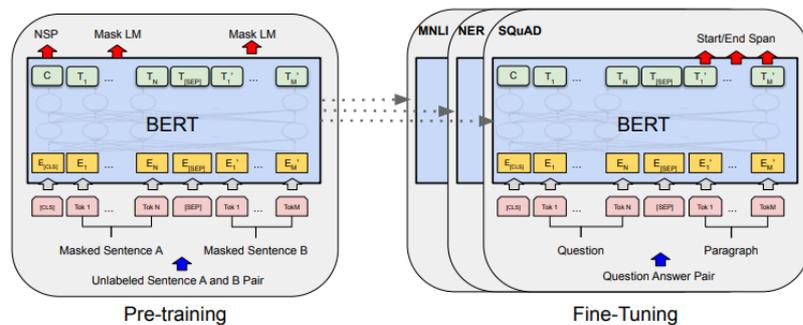
BERT dilatih dengan menggunakan dua metode utama: *masked language modeling* (MLM) dimana model akan memprediksi kata yang disembunyikan, dan *next sentence prediction* (NSP) untuk memprediksi kata selanjutnya [11]. Hal ini membuat model pralatih BERT hanya perlu melalui proses fine-tuning dengan

tambahanan satu layer luar saja untuk menghasilkan model yang berperforma tinggi dalam banyak tugas NLP [11].

2.3.2 Arsitektur Transformer Encoder pada BERT

Arsitektur dasar BERT dibangun di atas konsep *Transformer Encoder* yang diperkenalkan oleh Vaswani et al. (2017) [24]. BERT-base memiliki 12 lapisan *transformer encoder* sedangkan BERT-large memiliki 24 lapisan, masing-masing menggunakan fitur kunci yang disebut *self-attention* dan *multi-head attention*. Struktur ini memungkinkan BERT untuk memproses seluruh urutan masukan secara paralel, bukan secara sekuensial seperti model-model sebelumnya (misalnya RNN atau LSTM), sehingga dapat memahami konteks suatu kata dari keseluruhan kalimat secara lebih efektif [11, 24]. Arsitektur ini membuat BERT bekerja secara efektif dalam memetakan relasi antar-kata di sebuah kalimat tanpa perlu melihat batasan urutan sekuensial.

Gambaran umum arsitektur BERT, termasuk prosedur pre-training dan fine-tuning, dapat dilihat pada Gambar 2.3. Gambar ini menunjukkan bagaimana arsitektur Transformer Encoder yang sama digunakan dalam kedua fase tersebut, di mana model dilatih untuk memahami konteks bahasa secara bidirectional.



Gambar 2.3. Arsitektur BERT dalam prosedur pre-training dan fine-tuning

Sumber: [11]

Dalam model berbasis transformer terdapat mekanisme *self-attention*, dimana setiap token dalam sebuah kalimat menghitung tingkat "perhatian" terhadap setiap token lain dalam kalimat yang sama. Hal ini memungkinkan model untuk menangkap dependensi jarak jauh antar-kata, terlepas dari posisi fisiknya dalam urutan. Rumus perhitungan *self-attention* adalah sebagai berikut:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

Di mana komponen-komponen dalam Rumus 2.1 adalah sebagai berikut:

1. Query (Q): Matriks yang bertindak sebagai "penanya" untuk mencari relevansi terhadap semua token lain.
2. Key (K): Matriks yang berfungsi sebagai "kunci" yang akan dicocokkan dengan *Query* untuk menentukan tingkat kepentingan.
3. Value (V): Matriks yang membawa informasi aktual yang akan diambil dan dibobotkan berdasarkan skor perhatian.

Perhitungan skor perhatian (QK^T) dibagi dengan faktor skala $\sqrt{d_k}$ untuk menjaga stabilitas gradien selama pelatihan. Hasilnya kemudian dimasukkan ke dalam fungsi *softmax* untuk menghasilkan bobot perhatian α_{ij} yang nilainya berkisar antara 0 dan 1. Bobot inilah yang kemudian digunakan untuk membobotkan matriks *Value*, menghasilkan representasi akhir yang kaya akan konteks.

Selanjutnya, BERT mengadopsi *multi-head attention*, yang memungkinkan model untuk melakukan perhitungan *self-attention* secara paralel di berbagai ruang representasi yang berbeda (*heads*). Ini memberikan kemampuan pada model untuk menangkap berbagai jenis informasi kontekstual secara simultan. Rumus perhitungan *multi-head attention* adalah:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.2)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

Di mana h adalah jumlah *heads* atau lapisan (biasanya 12 untuk BERT-Base) dan W^O adalah matriks proyeksi akhir setelah hasil dari setiap *head* digabungkan (*concatenated*). Karena keunggulan teknologi transformer ini, BERT dapat melakukan perhitungan ke dua arah yang tidak dimiliki oleh model sekuensial lama seperti RNN/LSTM.

2.3.3 Representasi Input dan Tokenisasi pada BERT

Sebelum teks dimasukkan ke dalam model BERT, teks tersebut harus melalui proses *pre-processing* dan tokenisasi untuk diubah menjadi format numerik yang dapat dipahami oleh model. BERT menggunakan pendekatan WordPiece Tokenization (Wu et al., 2016), sebuah metode tokenisasi sub-kata yang memecah kata-kata menjadi unit-unit yang lebih kecil (sub-kata atau karakter) jika kata tersebut tidak ditemukan dalam kosakata model [11, 25]. Ini memungkinkan model untuk menangani kata-kata yang tidak dikenal (out-of-vocabulary) dan mengurangi ukuran kosakata yang diperlukan, sambil tetap mempertahankan informasi semantik.

Selain itu, BERT memerlukan penambahan token khusus dan representasi embedding tambahan untuk memahami struktur masukan:

1. Token Khusus:

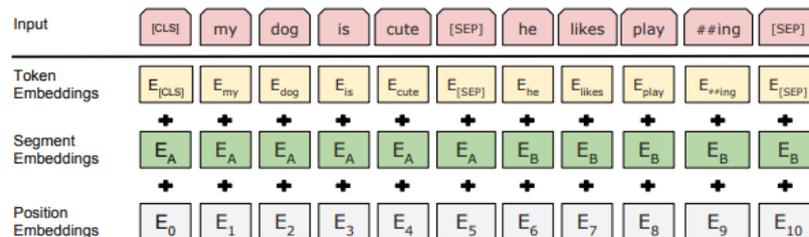
- (a) [CLS]: Ditambahkan di awal setiap urutan masukan. Vektor representasi akhir dari token ini sering digunakan sebagai representasi agregat untuk seluruh urutan, terutama untuk tugas klasifikasi [11].
- (b) [SEP]: Digunakan untuk memisahkan dua segmen kalimat (misalnya, pasangan kalimat A dan B dalam tugas *Next Sentence Prediction*). Untuk input tunggal, ini ditambahkan di akhir kalimat [11].

2. Input *Embeddings*: Representasi numerik untuk setiap token input dibentuk dari penjumlahan tiga jenis *embedding*[11]:

- (a) *Token Embeddings*: Representasi vektor dari setiap token (*WordPiece*).
- (b) *Segment Embeddings*: Menunjukkan segmen kalimat mana (A atau B) suatu token berasal. Digunakan untuk tugas yang melibatkan pasangan kalimat.
- (c) *Position Embeddings*: Memberikan informasi posisi token dalam urutan, karena *self-attention* itu sendiri tidak mempertimbangkan urutan.

Proses *pre-processing* ini juga mencakup langkah-langkah seperti normalisasi teks (mengubah semua teks menjadi huruf kecil untuk model *uncased*), pembersihan *whitespace*, dan penghapusan karakter yang tidak relevan, sebelum tokenisasi dilakukan. Hasilnya adalah urutan ID token yang siap untuk

dimasukkan ke dalam lapisan *encoder* BERT. Ilustrasi lengkap dari proses tokenisasi dan pembentukan representasi input BERT dapat dilihat pada Gambar 2.4.



Gambar 2.4. Representasi Input BERT

Sumber: [11]

2.3.4 IndoBERT: Adaptasi BERT untuk Bahasa Indonesia

IndoBERT adalah varian model BERT yang telah dilatih secara khusus dengan korpus bahasa Indonesia. Model ini memiliki *tokenizer* dan kosa kata Bahasa Indonesia sehingga dapat memproses struktur kalimat lokal. Hasilnya, model IndoBERT menunjukkan hasil yang sangat baik pada tugas - tugas NLU Bahasa Indonesia [26]. Penelitian yang dilakukan oleh Awalina et al. (2021) membuktikan bahwa BERT efektif dalam melakukan tugas klasifikasi berita hoaks dan mencapai hasil akurasi sebesar 90% [27]. Hal ini menegaskan bahwa model BERT yang telah di pralatih seperti IndoBERT sangat bermanfaat dalam mengerjakan berbagai tugas NLP Indonesia.

2.4 Pendekatan Deteksi Berita Hoaks Berbasis Teks

Deteksi berita hoaks berbasis teks merupakan salah satu bidang penelitian penting dalam upaya menjaga keamanan informasi dan siber. Berbagai pendekatan telah dikembangkan untuk mengidentifikasi konten palsu, yang secara umum dapat dikategorikan menjadi metode konvensional, pembelajaran *machine learning*, dan *deep learning* [28].

Secara tradisional, deteksi berita hoaks atau teks palsu dilakukan dengan teknik berbasis aturan atau statistika. Metode ini umumnya memeriksa ciri statistik teks seperti perplexity, stylometry (pola penulisan), dan frekuensi n-gram untuk mencari ketidakwajaran atau penyimpangan dari pola bahasa yang diharapkan [29].

Pada tingkat *machine learning*, pendekatan umum melibatkan ekstraksi fitur manual atau semi-otomatis dari teks, yang kemudian diklasifikasikan menggunakan algoritma seperti Support Vector Machine (SVM), Naive Bayes, atau Random Forest [30]. Fitur yang digunakan dapat mencakup fitur linguistik, fitur berbasis konten (misalnya kepadatan kata kunci), atau fitur berbasis kredibilitas sumber. Metode-metode ini menunjukkan peningkatan performa dibandingkan teknik konvensional.

Namun, seiring dengan kemajuan pesat AI, muncul teknik pendekatan baru yaitu *deep learning*. Perbandingan banyak studi menunjukkan bahwa model berbasis *deep learning*, khususnya yang menggunakan arsitektur transformer (seperti BERT), cenderung memiliki hasil yang lebih baik dalam mendeteksi berita palsu dibanding metode lain [31]. Kemampuan model *deep learning* untuk mempelajari representasi fitur yang kompleks secara otomatis dari data, serta memahami konteks semantik dan sintaksis yang rumit, menjadikannya sangat cocok untuk tugas deteksi berita hoaks. Tabel 2.1 merupakan contoh perkembangan pendekatan metode deteksi berita hoaks yang digunakan.

Tabel 2.1. Perkembangan Pendekatan Deteksi Berita Hoaks Berbasis Teks

Pendekatan Konvensional	Pendekatan <i>Machine learning</i>	Pendekatan <i>Deep learning</i>
Perplexity, stylometry (pola penulisan), dan frekuensi n-gram	Algoritma seperti Support Vector Machine (SVM), Naive Bayes, atau Random Forest	Arsitektur berbasis Transformer

2.5 Dataset untuk Deteksi Berita Hoaks Teks

Indonesian News Datasets adalah dataset yang berisikan artikel berita yang diambil dari tujuh website berita ternama di Indonesia yaitu: Tempo, CNN Indonesia, CNBC Indonesia, Okezone, Suara, Kumparan, and JawaPos. Dataset ini dapat ditemukan dan digunakan dari website Kaggle. Dataset ini terdiri dari 32.000 artikel berita dan memiliki sebelas kolom yaitu: id, source, title, image, url, content, date, embedding, created_at, updated_at, dan summary. TurnBackHoax.id merupakan website yang berisikan kumpulan artikel berita palsu (Hoaks). Website ini memiliki artikel berita hoaks mulai dari tahun 2016 hingga sekarang. Kumpulan - kumpulan artikel berita palsu akan didata setiap harinya dan di-*upload* ke website.

2.6 Studi Terkait

Penelitian dalam deteksi berita hoaks telah mengalami kemajuan signifikan dengan hadirnya model bahasa berbasis arsitektur Transformer [24], khususnya BERT (Bidirectional Encoder Representations from Transformers) [11]. Kemampuan BERT untuk memahami konteks kata secara dua arah menjadikannya sangat unggul untuk tugas klasifikasi teks. Di tingkat internasional, banyak penelitian telah mengadopsi BERT untuk membangun detektor berita palsu yang efektif. Berbagai penelitian telah membuktikan efektivitas model berbasis arsitektur BERT dalam mendeteksi berita hoaks dengan menyajikan hasil kuantitatif yang mengesankan. Dalam konteks internasional, penelitian oleh Kaliyar et al. [31] yang mengembangkan model BERT dengan CNN dalam melatih 20,800 dataset yang menghasilkan model FakeBERT dengan tingkat akurasi sebesar 98.90%.

Untuk konteks Bahasa Indonesia, penerapan model BERT yang dilatih secara lokal seperti IndoBERT juga menunjukkan hasil yang sangat menjanjikan. Penelitian yang dilakukan oleh Rahmawati et al. [30] secara spesifik menggunakan IndoBERT-base untuk klasifikasi berita hoaks dan berhasil memperoleh tingkat akurasi sebesar 90% dengan menggunakan 2,000 dataset. Studi lain oleh Fawaid et al. [27], yang membandingkan beberapa arsitektur (CNN, BiLSTM, Hybrid CNN-BiLSTM) termasuk Transformer (BERT), hasil penelitian mendemonstrasikan bahwa BERT mendapatkan hasil akurasi terbesar dengan nilai 90% pada 2,216 dataset berita berbahasa Indonesia. Capaian-capaian ini secara konsisten menunjukkan bahwa model berbasis Transformer merupakan state-of-the-art untuk deteksi hoaks dalam Bahasa Indonesia. Perbandingan dari hasil penelitian dapat dilihat pada Tabel 2.2.

Tabel 2.2. Perbandingan Hasil Akurasi dengan Penelitian Sebelumnya

Penelitian Terkait	Model	Jumlah Dataset	Akurasi
Kaliyar et al. (2021) [31]	FakeBERT	~20,800	98.90%
Rahmawati et al. (2022) [30]	IndoBERT-base	~2,000	90%
Fawaid et al. (2021) [27]	BERT	~2,216	90%