BULETIN PAGELARAN MAHASISWA NASIONAL BIDANG TEKNOLOGI INFORMASI DAN KOMUNIKASI
ISSN: xxxxxx

Identifikasi Pola Evolusi Modus Penipuan pada Forum Online berbasis Named Entity Recognition Menggunakan Indonesian Bidirectional Encoder Representation from Transformers

Moch Faiz Febriawan[1], Alisha Deana Tabina[2], Rakan Refaya Dewangga[3], Elly Matul Imah[4]

[1]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Surabaya, 60231 Indonesia, email: moch.faiz.23068@mhs.unesa.ac.id
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Surabaya, 60231 Indonesia, email: alishadeana.23244@mhs.unesa.ac.id
[3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Surabaya, 60231 Indonesia, email: rakan.23108@mhs.unesa.ac.id
[4]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Surabaya, 60231 Indonesia, email: ellymatul@unesa.ac.id
Corresponding Author: Elly Matul Imah

INTISARI — Penipuan daring di Indonesia terus berkembang dengan berbagai modus yang memanfaatkan media sosial dan forum online. Penelitian ini bertujuan mengidentifikasi evolusi modus penipuan melalui ekstraksi entitas penting menggunakan metode Named Entity Recognition (NER) berbasis IndoBERT. Data dikumpulkan dari X (Twitter), Reddit, dan Quora selama 2020 hingga pertengahan 2025. Model IndoBERT digunakan untuk mendeteksi entitas seperti modus penipuan, produk, layanan, nominal uang, dan platform. Entitas yang terdeteksi diklasifikasikan ke dalam enam kategori modus menggunakan algoritma Logistic Regression. Model NER menunjukkan akurasi 96% dengan metrik evaluasi yang kuat, sementara model klasifikasi mencapai akurasi 70,67%. Analisis menunjukkan dominasi modus *Social Engineering* melalui aplikasi pesan dan media sosial, sedangkan analisis temporal memperlihatkan lonjakan kasus modus donasi palsu pada pertengahan tahun 2024, serta tren kenaikan modus keuangan dan produk digital sepanjang periode penelitian. Pendekatan gabungan NER berbasis IndoBERT dan klasifikasi teks ini terbukti efektif untuk deteksi dini dan analisis modus penipuan daring di Indonesia, sehingga dapat mendukung upaya pencegahan, mitigasi risiko, serta meningkatkan literasi digital masyarakat dalam menghadapi ancaman penipuan daring.

KATA KUNCI — Evolusi modus penipuan, Named Entity Recognition (NER), Indonesian Bidirectional Encoder Representation from Transformers (IndoBERT), klasifikasi modus penipuan, machine learning

I. PENDAHULUAN

A. LATAR BELAKANG

Indonesia mengalami peningkatan signifikan kasus penipuan daring, dengan 166.258 laporan diterima oleh OJK dan IASC antara November 2024 hingga Juni 2025. Modus penipuan mencakup pinjaman online ilegal, investasi palsu, donasi, dan lainnya yang marak di media sosial dan forum online. Penyebaran modus ini memanfaatkan media digital untuk memanipulasi korban, sehingga penting untuk mendeteksi dan mengklasifikasikan jenis penipuan serta pola bahasanya. Upaya ini tidak hanya meningkatkan literasi digital, tetapi juga mendukung SDGs, khususnya poin 1 (Tanpa Kemiskinan), 10 (Mengurangi Kesenjangan), dan 16 (Perdamaian, Keadilan, dan Lembaga yang Kuat) .

Beberapa penelitian terdahulu telah menerapkan algoritma *machine learning* sebagai pendekatan dalam identifikasi penipuan di media sosial. Mitrakanasih Laia dan tim menggabungkan Big Data, NLP, dan ML untuk meningkatkan akurasi deteksi hoaks[1]. Algoritma seperti BI-LSTM, SVM, Random Forest, dan Decision Tree juga digunakan pada platform seperti X (dulu Twitter)[2][3]. Yajing Liu dkk. mengembangkan HA-GNN untuk menangani kamuflase identitas dalam jaringan sosial [4]. Arif Ridho Lubis menerapkan deep learning untuk mendeteksi akun penipuan, sedangkan Najla Arhabi menggunakan LSTM untuk klasifikasi akun palsu di Instagram [5][6].

Penelitian lain juga memanfaatkan kombinasi model canggih seperti YOLOv8, PaddleOCR, dan Fuzzy Matching untuk deteksi penipuan dalam penggalangan dana medis[7]. Lukman Hakim menggunakan IndoBERT untuk membangun NER dalam mendeteksi aktivitas kriminal pada artikel berita[8]. Sandeep Pamarthi menerapkan NER untuk identifikasi dan anonimisasi entitas dalam data transaksi keuangan[9].

Beberapa studi menggunakan dataset hasil scraping media sosial. Hansen Dusenov menganalisis cuitan berbahasa Indonesia terkait penipuan, sementara Acharya dan rekan-rekannya mempelajari lebih dari 151.000 akun dan 3 juta unggahan di X, Instagram, Facebook, YouTube, dan Telegram (Maret–Mei 2024), termasuk modus website fundraising palsu dan tautan eksternal menyesatkan[3][10].

Dalam penelitian ini, data dikumpulkan dari media sosial X dan forum diskusi seperti Reddit dan Quora. Metode Named Entity Recognition (NER) digunakan untuk mengekstrak entitas penting seperti pelaku, jenis modus, dan platform yang digunakan. Untuk meningkatkan akurasi, digunakan model IndoBERT yang telah dilatih pada korpus bahasa Indonesia. Berdasarkan penelitian Syaiful Imron, IndoBERT terbukti unggul dalam memahami konteks, bahkan mengungguli M-BERT dalam tugas berbahasa Indonesia[11]. Oleh karena itu, penelitian ini mengembangkan model NER berbasis IndoBERT untuk mendeteksi modus penipuan dari pengalaman pengguna secara daring.

B. TUJUAN

Penipuan digital saat ini menjadi hal krusial yang terus berkembang, terutama di media sosial dan forum online tempat manusia berinteraksi serta berdiskusi secara terbuka. Dalam

ISSN: xxxxxx

konteks ini, tujuan analisis ini adalah untuk menggali pemahaman yang lebih mendalam mengenai pola modus penipuan melalui pengalaman pengguna di laman pribadinya. analisis ini berupaya mengidentifikasi tren dan evolusi modus penipuan dari waktu ke waktu.

Harapannya penelitian ini mampu memberikan wawasan mengenai bagaimana modus penipuan berevolusi dari waktu ke waktu, sekaligus mengungkap ciri-ciri komunikasi para pelaku secara lebih mendetail.

C. MANFAAT

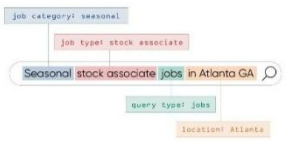
Pendekatan ini juga memungkinkan deteksi dini terhadap modus-modus baru yang digunakan oleh pelaku, sehingga upaya pencegahan dapat dilakukan secara lebih efektif. Selain itu, pihak-pihak yang berwenang seperti lembaga pengawas, aparat penegak hukum, maupun platform media sosial dapat merumuskan solusi yang lebih adaptif dan responsif dalam menghadapi perkembangan modus penipuan. Dengan demikian, jumlah korban penipuan di media sosial dapat ditekan secara signifikan, dan ekosistem digital menjadi lebih aman bagi seluruh pengguna.

D. BATASAN DAN RUANG LINGKUP

Identifikasi pola modus penipuan dilakukan dengan scraping data dari forum online seperti Twitter, Reddit, dan Quora, terbatas pada unggahan berbahasa Indonesia dengan kurun waktu awal tahun 2020 hingga pertengahan tahun 2025. Sistem menggunakan metode *Named Entity Recognition* (NER) berbasis IndoBERT untuk mengekstraksi entitas seperti jenis modus, pelaku, tindakan, informasi finansial, dan platform. Klasifikasi dibatasi pada kategori modus tertentu (produk digital, fisik, keuangan) serta vektor serangan seperti phishing dan malware. Analisis difokuskan pada pola bahasa dan evolusi modus penipuan, tanpa mencakup pelacakan individu atau tindakan hukum.

II. STUDI LITERATUR

A. NAMED ENTITY RECOGNITION



Gambar 1. Ilustrasi Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah salah satu sub-tugas dalam *text mining* dan *Natural Language Processing (NLP)* yang bertujuan untuk mengidentifikasi dan mengelompokkan entitas penting dalam sebuah teks ke dalam kategori tertentu[12]. NER berperan besar dalam mengekstraksi informasi dari data teks yang tidak terstruktur, sehingga sangat berguna untuk berbagai aplikasi seperti ekstraksi informasi dari artikel berita, analisis media sosial, chatbot, dan sistem deteksi penipuan [13]. Dalam konteks penelitian ini, NER digunakan untuk menandai entitas penting yang berkaitan dengan penipuan, seperti nama pelaku, modus penipuan, aksi penipuan, nominal uang, layanan, produk, rekening, kontak, dan platform tempat penipuan terjadi.

Pendekatan NER dibagi menjadi pendekatan berbasis aturan dan pendekatan pembelajaran mesin, yang mencakup model-statistik klasik hingga model transformer seperti IndoBERT. Pendekatan Named Entity Recognition (NER) berbasis aturan (rule-based) mengandalkan aturan linguistik dan kamus yang disusun secara manual untuk mengidentifikasi entitas dalam teks [14]. Sementara itu, pendekatan berbasis statistik dan pembelajaran mesin (*machine learning*) menggunakan algoritma seperti *Conditional Random Fields* (CRF), *Hidden Markov Models* (HMM), hingga metode deep learning terkini seperti *Recurrent Neural Networks* (RNN) serta model berbasis Transformer seperti BERT dan IndoBERT sangat efektif dalam menangani konteks dan variasi data yang kompleks [15].

Tantangan utama dalam pengembangan NER bahasa Indonesia terletak pada variasi dialek, ambiguitas morfologi, dan data anotasi yang terbatas di domain tertentu [16]. Oleh karena itu, salah satu perkembangan NER adalah menggunakan Transformer-based models seperti BERT (*Bidirectional Encoder Representations from Transformers*), yang mampu memahami konteks kata secara lebih baik melalui mekanisme *self-attention* [17]. Penelitian ini menggunakan IndoBERT, model bahasa berbasis BERT yang dilatih secara khusus menggunakan data dalam bahasa Indonesia.

B. INDONESIAN BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMER

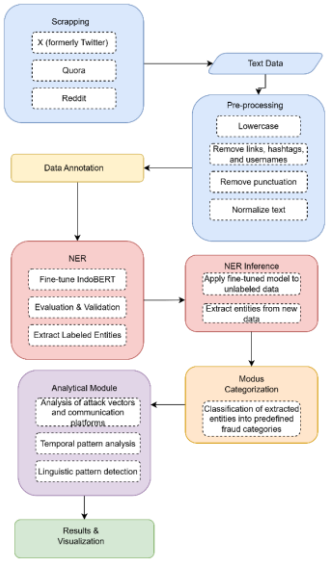
IndoBERT adalah model bahasa berbasis transformer yang dikembangkan secara khusus untuk menangani keunikan linguistik dan ragam dialek dalam bahasa Indonesia, sehingga mampu menunjang berbagai tugas NLP secara lebih akurat dibandingkan model multibahasa seperti mBERT[18]. Jika dibandingkan dengan model multibahasa seperti mBERT, model ini dapat memberikan hasil yang lebih akurat karena didasarkan pada korpus bahasa Indonesia yang besar [19]. Dalam berbagai studi, IndoBERT menunjukkan performa yang unggul untuk berbagai tugas NLP seperti klasifikasi teks, analisis sentimen, serta Pengenalan Entitas Bernama (*Named Entity Recognition*/NER)[20]. Salah satu kekuatan utamanya adalah kemampuan memahami hubungan antar-token dalam konteks kalimat, yang didukung oleh mekanisme *self-attention* dalam arsitektur transformer [17][21]. Skema *self-attention* dapat dituliskan secara matematis pada persamaan (1)

$$Attention(Q,K,V) = softmax\left(\frac{Qk^T}{\sqrt{dk}}\right)V \qquad (1)$$

Setiap token dalam kalimat akan diproses melalui lapisan encoder dengan mekanisme *self-attention* untuk menghasilkan representasi kontekstual yang digunakan dalam berbagai tugas NLP seperti NER. Dalam konteks *Named Entity Recognition* (NER), IndoBERT digunakan untuk mengklasifikasikan setiap token dalam kalimat ke dalam label entitas tertentu seperti *PERSON*, *LOCATION*, atau *ORGANIZATION*. Kemampuan representasi kontekstual IndoBERT membantu meningkatkan akurasi ekstraksi entitas dalam bahasa Indonesia yang memiliki struktur dan morfologi yang kompleks

III. USULAN SOLUSI

Sistem yang diusulkan bertujuan untuk mendeteksi dan mengidentifikasi modus penipuan dari berbagai forum online populer seperti Twitter, Reddit, dan Quora. Sistem ini menggunakan pendekatan pemrosesan bahasa alami berbasis *Named Entity Recognition* (NER) dengan model bahasa Indonesia IndoBERT. Arsitektur sistem secara umum dapat dilihat pada gambar 2.



Gambar 2. Diagram Arsitektur Sistem

A. DATA DAN PRA-PEMROSESAN DATA

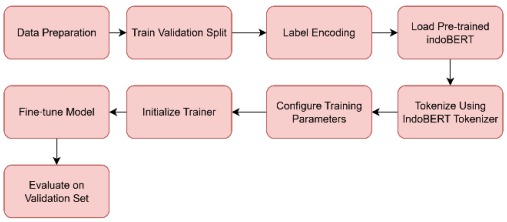
Data yang digunakan dalam penelitian deteksi penipuan berbasis teks ini didapatkan dari tiga platfrom daring, yaitu media sosial X, Quora, dan Reddit. Total data yang berhasil terkumpul adalah sebanyak 4061 entri yang mengandung indikasi atau narasi pengalaman pengguna terkait tindakan penipuan. Pada proses ini dilakukan pengumpulan data dengan menggunakan metode scraping yang mengambil postingan berkaitan dengan penipuan. Proses scraping dilakukan menggunakan API masing-masing platform dengan kata kunci seperti "penipuan", "modus", "ketipu", dan "scam". Tentunya, setiap entri data berbahasa indonesia dan memuat tanggal dari waktu posting. Setelah seluruh data berhasil digabungkan, tahap selanjutnya adalah pra-pemrosesan data. Langkah ini mencakup, mengubah huruf menjadi huruf kecil (lowercase), menghapus tautan, hashtag, dan mention, menghapus tanda baca, penghapusan data duplikat, serta melakukan normalisasi kata tidak baku.

B. ANOTASI DATA

Proses ini dilakukan dengan pendekatan semi-otomatis. *Generative AI* digunakan untuk memberikan label awal terhadap data teks berdasarkan skema *Named Entity Recognition* (NER) yang telah ditentukan, yaitu meliputi BM/IM (Modus), BP/IP (Pelaku), BR/IR (Rekening), BA/IA (Aksi), BN/IN (Nominal), B-KONTAK/I-KONTAK (Kontak), B-PRODUK/I-PRODUK (Produk), B-LAYANAN/I-LAYANAN (Layanan), dan Platform. Skema pelabelan yang digunakan adalah BIO (Beginning, Inside, Outside), karena skema ini memberikan kejelasan posisi token dalam suatu entitas serta terbukti efektif dalam berbagai tugas sequence labeling. Label yang dihasilkan oleh AI kemudian diperiksa dan divalidasi oleh peneliti guna memastikan akurasi dan konsistensi anotasi. Pendekatan ini dipilih untuk mempercepat proses anotasi data dalam jumlah besar sekaligus menjaga kualitas hasil anotasi.

C. PEMODELAN DENGAN INDOBERT

Untuk mengidentifikasi entitas-entitas penting dalam teks naratif berbasis penipuan, pemodelan dilakukan menggunakan pendekatan *sequence labeling*. *Sequence Labeling* adalah tugas mendasar dalam pemrosesan bahasa alami ini melibatkan pemberian label pada setiap kata atau elemen dalam sebuah kalimat, dengan tujuan untuk mengidentifikasi fungsi atau kategorinya, seperti nama orang, tindakan, atau objek tertentu (Xinyu wang). Pemodelan dilakukan untuk mengidentifikasi entitas-entitas penting dalam teks naratif berbasis penipuan menggunakan pendekatan *sequence labeling*. Model yang digunakan adalah IndoBERT. IndoBERT dipilih karena telah terbukti efektif dalam memahami konteks kalimat dalam bahasa Indonesia. Model ini dioptimalkan untuk berbagai tugas NLP seperti *named entity recognition (NER)*, sehingga sesuai untuk digunakan dalam penelitian ini. Adapun tahap pemodelannya dapat dilihat pada gambar 3.



Gambar 3. Alur Pemodelan NER

Proses pelatihan model NER berbasis IndoBERT dimulai dengan menyiapkan data dalam format token dan label menggunakan skema BIO. Setiap token diberi label seperti B-MODUS, I-MODUS, dan lain-lain, sementara baris kosong digunakan untuk memisahkan antar kalimat. Hasilnya, setiap kalimat dikonversi menjadi daftar token dan daftar label yang cocok. Setelah data siap, dilakukan pembagian menjadi data latih dengan proprosi 90% dan data validasi 10%. Selanjutnya, dilakukan label encoding yaitu mapping label-ke-ID dan ID-ke-label. Model IndoBERT pre-trained dari HuggingFace, dengan arsitektur 12-layer dan 110 juta parameter, digunakan sebagai dasar. Model ini kemudian ditambahkan linear layer di bagian akhir untuk mengklasifikasikan setiap token ke dalam label NER yang sesuai. Tokenisasi dilakukan menggunakan tokenizer IndoBERT berbasis WordPiece. Model dilatih selama 5 epoch dengan batch size 16 per device menggunakan optimizer AdamW. Skema pembelajaran menggunakan cosine learning rate decay, warmup ratio sebesar 0.1, dan weight decay 0.01 untuk mengurangi risiko overfitting. Model dan log hasil pelatihan disimpan di direktori ./ner-indobert dan ./logs.

Setelah itu, proses pelatihan dibantu oleh modul Trainer dari HuggingFace, yang mempermudah semuanya. *Fine-tuning* dilakukan secara penuh, artinya seluruh parameter model dengan total sebanyak 123.865.363 diperbarui selama proses pelatihan agar lebih sesuai dengan karakteristik data. Evaluasi performa model dilakukan menggunakan metrik presisi, recall, dan F1-score untuk menilai akurasi dalam mendeteksi entitas secara keseluruhan. Tahap evaluasi model menggunakan evaluasi *precision* (presisi), *recall* (sensitivitas), dan F1-score. Ketiga metrik ini sangat umum digunakan dalam evaluasi model klasifikasi karena mampu menggambarkan seberapa baik model mengenali dan mengklasifikasikan entitas dengan benar. Perhitungan ketiga metrik ini didasarkan pada jumlah *True Positive* (TP), yaitu entitas yang diprediksi benar oleh model; *False Positive* (FP), yaitu entitas yang diprediksi oleh model tetapi sebenarnya salah; dan *False Negative* (FN), yaitu entitas yang seharusnya dikenali namun tidak berhasil diprediksi oleh model. Rumus matriks evalusi matriks seperti persamaan (2), (3), dan (4)

$$precision = \frac{TP}{TP+FP} \quad (2)$$

ISSN: xxxxxx

$$recall = \frac{TP}{TP+FN} \quad (3)$$

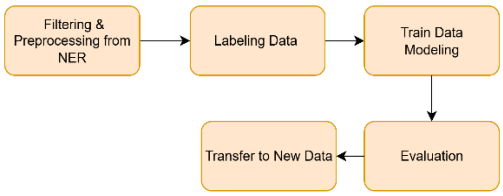
$$F1 = 2 \cdot \frac{(precision \cdot recall)}{precision+recall} \quad (4)$$

D. PENERAPAN MODEL KE DATA BARU

Setelah model berhasil dilatih dan divalidasi, langkah selanjutnya adalah menerapkan model tersebut pada data baru yang belum teranotasi. Proses ini bertujuan untuk mengidentifikasi entitas-entitas penting secara otomatis, seperti modus penipuan, layanan/produk, jumlah nominal, rekening tujuan, serta aksi yang dilakukan oleh pelaku penipuan. Proses ini memanfaatkan pendekatan *transfer learning*, di mana model yang telah di*fine-tune* dapat langsung digunakan untuk memproses data baru tanpa perlu pelatihan ulang.

E. KLASIFIKASI DAN ANALISIS POLA PENIPUAN

Data yang telah dianotasi dengan NER, setiap entitas yang berhasil diekstraksi akan dipetakan ke dalam kategori modus penipuan, yaitu Produk Digital, Produk Fisik, Keuangan dan Investasi, Jasa online dan Ketenagakerjaan, Properti dan Akomodasi, serta Kepercayaan Sosial dan donasi. Klasifikasi ini dilakukan dengan algoritma *machine learning*. Proses klasifikasi dapat dilihat pada gambar 4.



Gambar 4. Alur Klasifikasi Kategori Penipuan

Penelitian ini menggunakan 2.072 entri data yang diperoleh melalui proses Named Entity Recognition (NER). Data tersebut kemudian disaring dan diproses sebagai bagian dari tahapan pra-pemrosesan. Untuk mempercepat proses anotasi, sekitar 1.500 entri dilabeli menggunakan bantuan generative AI. Hasil pelabelan ini kemudian ditinjau ulang secara manual oleh peneliti guna memastikan validitas dan ketepatan klasifikasi data. Setelah proses pelabelan selesai, data digunakan untuk melatih model klasifikasi dengan pendekatan machine learning, menggunakan algoritma Logistic Regression. Evaluasi dilakukan untuk menilai performa model melalui metrik seperti precision, recall, dan F1-score, yang masing-masing dihitung menggunakan rumus pada persamaan (2), (3), (4), dan (5)

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \times 100\% \quad (5)$$

Model terbaik ini kemudian digunakan dalam proses klasifikasi lanjutan terhadap data baru sebanyak 1.838 entri, menggunakan pendekatan transfer learning. Dengan metode ini, model mampu mengenali pola pada data yang belum pernah dilihat sebelumnya dengan mengandalkan pembelajaran dari data terdahulu.

F. ANALISIS

Setelah proses klasifikasi selesai, data akan digunakan untuk analisis lanjutan untuk mendapatkan wawasan mengenai pola-pola penipuan. Analisis pertama yang dilakukan adalah analisis Vektor Serangan dan Kanal dengan tujuan mengidentifikasi hubungan antara jenis modus penipuan (vektor serangan) dan kanal atau platform komunikasi yang digunakan pelaku. Vektor serangan yang diidentifikasi dikelompokkan dalam penelitian ini, yaitu *Phising*, Malware/APK, *Impersonation*, dan *Social Engineering.* Selanjutnya dilakukan analisis untuk mendeteksi pola bahasa dengan mengevaluasi pola-pola bahasa yang digunakan pelaku penipuan dengan pendekatan *n-gram*, yaitu teknik yang memecah teks menjadi sekuens kata atau frasa (misalnya unigram dan bigram) untuk mengidentifikasi kemunculan kata atau frasa yang sering digunakan [22]. Analisis terakhir adalah analisis temporal, yaitu menganalisis pola waktu terjadinya penipuan, misalnya tren modus tertentu yang meningkat pada bulan-bulan tertentu, atau intensitas aktivitas penipuan yang meningkat selama periode tertentu.

IV. HASIL DAN PEMBAHASAN

A. HASIL ANOTASI DATA

Dari total 4050 data teks, data yang teranotasi secara emi-otomatis adalah sebanyak 1056 data berhasil dianotasi menggunakan skema label BIO (Beginning, Inside, Outside). Adapun label entitas yang digunakan dalam anotasi mengikuti skema sebagai berikut

- MODUS (B-MODUS, I-MODUS): Menandakan jenis penipuan.
- PERSON (B-PERSON, I-PERSON): Pelaku atau identitas manusia.
- REK (B-REK, I-REK): Informasi rekening.
- ACTION (B-ACTION, I-ACTION): Tindakan atau instruksi penipuan.
- NOMINAL (B-NOMINAL, I-NOMINAL): Nilai atau jumlah uang.
- PRODUK (B-PRODUK, I-PRODUK): Produk atau layanan yang digunakan sebagai umpan.
- KONTAK (B-KONTAK, I-KONTAK): Informasi kontak seperti nomor HP atau email.

Data dari hasil anotasi ini kemudian digunakan sebagai data latih (*training data)* untuk membangun model *Named Entity Recognition (*NER*)* berbasis IndoBERT.

B. EVALUASI MODEL NER INDOBERT

Evaluasi dilakukan menggunakan *metrik Precision*, *Recall*, dan *F1-Score* per label untuk mengukur kinerja model dalam mendeteksi setiap entitas. Hasil evaluasi model dapat dilihat pada Tabel berikut

Tabel 1. Evaluasi Model NER IndoBERT

Label Entitas	Precision	Recall	F1-Score	Support
ACTION	0.95	0.95	0.95	151
KONTAK	0.73	0.8	0.76	10
LAYANAN	0.54	0.62	0.58	21
MODUS	0.88	0.88	0.88	165
NOMINAL	0.81	0.85	0.83	26
PERSON	0.73	0.69	0.71	32
PLATFORM	0.82	0.85	0.84	27
PRODUK	0.59	0.56	0.57	54
REK	0.89	1	0.94	8
accuracy			0.96	494
micro avg	0.84	0.84	0.84	494
macro avg	0.77	0.8	0.78	494
weighted avg	0.83	0.84	0.84	494

Berdasarkan hasil evaluasi pada Tabel 1, model IndoBERT yang telah di *fine-tuned* menunjukkan performa yang sangat baik dengan akurasi keseluruhan sebesar 96%. Label-label seperti *ACTION*, *MODUS*, dan *REK* memperoleh skor precision dan recall yang tinggi, menunjukkan bahwa model mampu mengenali entitas-entitas tersebut secara konsisten. Namun, terdapat beberapa label seperti *LAYANAN* dan *PRODUK* yang memiliki nilai f1-score relatif lebih rendah, mengindikasikan tantangan dalam mengenali entitas tersebut,

kemungkinan akibat jumlah data yang lebih sedikit karena pengguna tidak menyebutkan tentang produk dan layanan saat membagikan pengalaman penipuannya atau adanya variasi penulisan dalam entitas tersebut.

C. TRANSFER LEARNING NER

Setelah model IndoBERT berhasil di fine-tuned dan menghasilkan metrik evaluasi yang baik, maka model yang sudah dibangun di terapkan di sisa data yang berjumlah 3005 data. Dari jumlah tersebut, sebanyak 2072 data berhasil diprediksi entitasnya secara otomatis, sementara sisanya tidak dapat diproses karena tidak mengandung pola teks yang sesuai atau entitas yang dapat dikenali oleh model.

Model yang telah melalui proses transfer learning ini mampu mengenali berbagai entitas seperti modul penipuan (MODUS), pelaku (PERSON), nomor rekening (REKENING), aksi (ACTION), nominal (NOMINAL), kontak, hingga nama produk atau layanan yang digunakan dalam modus penipuan. Berikut beberapa contoh hasil prediksi entitas dari NER

Tabel 2. Contoh hasil prediksi NER

Teks	Entitas yang Terdeteksi
hampir kena scam "skema segitiga" jual beli mobil	MODUS: scam skema segitiga, LAYANAN: jual beli, PRODUK: mobil
ortu gw kena hack akun dana, duit hilang jutaan	ACTION: hack, PLATFORM: dana, NOMINAL: jutaan
anggota keluarga kena scam "cashback donasi bantu anak" puluhan juta. apk network logs inside	MODUS: scam, donasi, ACTION: bantu anak, NOMINAL: puluhan juta
komodos yang pernah ketipu perkara jual beli tanah atau properti, what's your story?	ACTION: ketipu, LAYANAN: jual beli, PRODUK: tanah properti

Prediksi ini menunjukkan bahwa model mampu mengenali pola-pola entitas dalam kalimat-kalimat tidak terstruktur sekaligus, yang sangat berguna dalam proses identifikasi pola penipuan secara otomatis.

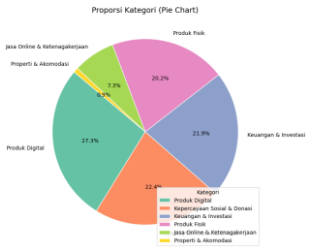
D. KLASIFIKASI

Pada tahap ini dilakukan klasifikasi data hasil ekstraksi entitas menggunakan model machine learning untuk mengelompokkan modus penipuan ke dalam enam kategori: Produk Digital, Produk Fisik, Keuangan & Investasi, Jasa Online & Ketenagakerjaan, Properti & Akomodasi, serta Kepercayaan Sosial & Donasi. Dataset yang digunakan untuk pelatihan model terdiri dari 1.500 data yang telah dilabeli, kemudian dilakukan pembagian data untuk pelatihan dan pengujian model. Model klasifikasi teks yang dibangun menggunakan algoritma Logistic Regression berhasil mencapai akurasi sebesar 70,67% pada data uji sebanyak 300 data. Hasil evaluasi kinerja model disajikan pada tabel berikut

Tabel 3. Evaluasi Klasifikasi Modus Penipuan

Kategori	Precision	Recall	F1-Score	Support
Jasa Online & Ketenagakerjaan	0.67	0.5	0.57	8
Kepercayaan Sosial & Donasi	0.81	0.88	0.84	72
Keuangan & Investasi	0.66	0.49	0.56	72
Produk Digital	0.63	0.81	0.71	72
Produk Fisik	0.73	0.71	0.72	72
Properti & Akomodasi	1	0.25	0.4	4
accuracy			0.71	300
macro avg	0.75	0.6	0.63	300
weighted avg	0.71	0.71	0.7	300

Setelah model dikembangkan dan diuji, model ini kemudian digunakan untuk mengklasifikasikan sisa data sebanyak 1.838 data yang belum diberi label secara manual. Berikut adalah distribusi kategori berdasarkan hasil klasifikasi terhadap seluruh data



Gambar 5. Proporsi Klasifikasi Kategorisasi

Temuan ini menyatakan bahwasannya Produk Digital serta Kepercayaan Sosial dan Donasi menjadi modus penipuan paling dominan yang sering di alami para pengguna.

E. ANALISIS LANJUTAN

1) Analisis vektor serangan dan kanal

Analisis ini bertujuan untuk mengidentifikasi keterkaitan antara jenis modus penipuan (vektor serangan) dan kanal komunikasi yang digunakan oleh pelaku. Berdasarkan hasil klasifikasi dan menggunakan *Named Entity Recognition* (NER), dari total data yang dianalisis, hanya 1.167 data yang berhasil diidentifikasi memiliki entitas yang dapat diolah kembali menjadi vektor serangan dan kanal. Temuan ini menunjukkan bahwa pelaku penipuan cenderung memilih kanal yang sesuai dengan strategi serangan mereka untuk meningkatkan efektivitas penipuan.

Berdasarkan hasil analisis lanjutan, *Social Engineering* merupakan vektor serangan yang paling dominan, terutama ketika dilakukan melalui *Messaging App* (449 kasus) dan *Social Media* (370 kasus), dan disusul Impersonation melalui Social Media (84 kasus). Modus ini memanfaatkan teknik manipulasi psikologis, tanpa memerlukan teknologi yang kompleks, biasanya pelaku memanfaatkan kedekatan pribadi yang disediakan oleh aplikasi pesan atau media sosial. Berikut tiga kombinasi modus penipuan dan kanal komunikasi yang paling sering ditemukan adalah

Tabel 4. 1Tiga Kombinasi Modus Penipuan dan Kanal Komunikasi Paling Sering Terjadi

No.	Vektor Serangan	Kanal	Contoh Kalimat
1	Social Engineering	Messaging App	“ati hati udah sampe bikin grup whatsapp penipuannya. 140 an member lagi semoga gaada yg ketipu :(“
2	Social Engineering	Social Media	"guys aku minta tolong buat report akun instagram /monseur.store dong dia udah curi pict aku pliss bantu aku dia ini scam ngambil foto orang lain terus dijual di akunnya. "
3	Impersonation	Social Media	"share pengalaman hampir tertipu, penipu yang mengatasnamakan shopee express."

Berdasarkan tabel tersebut, modus *Social Engineering* menjadi yang paling dominan, terutama melalui *Messaging App* dan *Social Media*, menandakan bahwa pelaku memanfaatkan komunikasi yang bersifat personal dan cepat untuk menipu

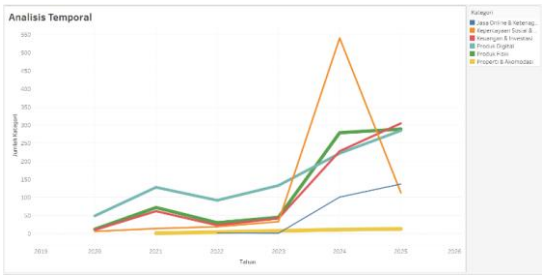
ISSN: xxxxxx

korban secara langsung. Berdasarkan hasil ekstraksi fitur dengan pendekatan n-gram (unigram dan bigram), berikut ini adalah kata-kata atau frasa yang paling sering muncul pada kategori Impersonation:

Tabel 5. Pola kata per-kategori yang sering digunakan

Kategori	Pola Bahasa Umum dan Kata Kunci Dominan
Social Engineering	Kata dengan muatan emosional tinggi seperti: *penipuan, ketipu, hati, uang, modus, ga, aja, banyak* : menunjukkan narasi personal dan ajakan hati-hati.
Impersonation	Ciri khas: penggunaan kata formal dan identitas palsu seperti *mengatasnamakan, mohon, kak, nomor* : menggambarkan peniruan institusi resmi.
Phishing	Kata seperti *link, langsung, buat, nomor, tersebut* menunjukkan ajakan untuk mengklik atau mengisi sesuatu secara cepat.
Malware/APK	Terlihat informal dan akrab, seperti *ibu, teman, bilang, langsung* : menandakan modus pengiriman file dari orang dekat atau keluarga.

2) Analisis Temporal



Gambar 6. Evolusi Modus Penipuan dari Waktu ke Waktu

Dari hasil analisis temporal yang ditampilkan pada Gambar 6, terlihat adanya peningkatan signifikan pada modus Kepercayaan Sosial & Donasi pada pertengahan tahun 2024. Fenomena ini kemungkinan besar dipengaruhi oleh situasi aktual di Indonesia, di mana BNPB mencatat 122 kejadian bencana alam selama periode tersebut. Peristiwa tersebut memicu maraknya penggalangan dana sosial dan kampanye amal daring yang rawan dimanfaatkan oleh oknum tidak bertanggung jawab sebagai sarana penipuan. Selain itu, modus Produk Digital serta Jasa Online & Ketenagakerjaan menunjukkan tren peningkatan secara bertahap sepanjang periode 2020 hingga 2025. Hal ini mencerminkan meningkatnya jumlah korban seiring dengan berkembangnya aktivitas digital masyarakat.

Puncak aktivitas modus penipuan juga cenderung bersifat musiman, terutama saat momen promosi besar seperti Hari Belanja Online Nasional (Harbolnas) dan menjelang akhir tahun, di mana terjadi lonjakan transaksi digital dan interaksi di media sosial. Secara keseluruhan, pola evolusi modus penipuan tidak hanya mencerminkan tren global, tetapi juga dipengaruhi oleh konteks sosial dan ekonomi lokal di Indonesia. Oleh karena itu, strategi pencegahan penipuan harus dirancang secara adaptif dan kontekstual, menyesuaikan dengan dinamika lokal yang terjadi.

V. KESIMPULAN

Penelitian ini menganalisis pengalaman masyarakat terhadap berbagai modus penipuan secara daring dengan memanfaatkan data teks berbahasa Indonesia yang dikumpulkan dari berbagai platform media sosial. Untuk mencapai tujuan tersebut, penelitian ini menggabungkan teknik *Named Entity Recognition* (NER) berbasis IndoBERT dan klasifikasi teks. Model ini memungkinkan sistem untuk mengenali dan mengelompokkan entitas penting dalam teks berbahasa Indonesia, seperti modus, pelaku, rekening, aksi, nominal, kontak, dan produk.

Hasil ekstraksi entitas selanjutnya digunakan untuk mengklasifikasikan teks ke dalam enam kategori penipuan beserta aksi dan kanal yang digunakan pelaku. Model klasifikasi yang dibangun menunjukkan performa yang baik dalam mengenali jenis penipuan berdasarkan informasi entitas yang tersedia. Hasil klasifikasi menunjukkan bahwa modus penipuan yang paling dominan adalah impersonasi dan penipuan berkedok pekerjaan, sementara kanal penipuan yang paling sering digunakan adalah aplikasi pesan instan dan media sosial. Dari analisis yang dilakukan, dapat disimpulkan bahwa modus penipuan semakin beragam dan terstruktur, dengan pola penyebaran yang sistematis melalui kanal komunikasi yang mudah diakses oleh masyarakat. Pendekatan gabungan ini terbukti efektif untuk deteksi awal penipuan daring dan berpotensi dikembangkan sebagai alat bantu keamanan digital di Indonesia.

REFERENSI

[1] M. Laia, Ayuliana, Wasiran, M. L. Hakim, and D. Suryadi, "Analisis Big Data untuk Deteksi Hoaks dan Disinformasi di Platform Berita Online," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 9, no. 2, pp. 776–782, 2025, doi: 10.35870/jtik.v9i2.3859.

[2] M. F. Alfawaid and J. Prianggono, "Penerapan Metode Multiple Machine Learning (Hybrid Model) Untuk Mendeteksi Link Phising Sebagai Upaya Preventif Dalam Meminimalisir Korban Pencurian Data," *J. Ilmu Kepol.*, vol. 17, no. 3, 2023.

[3] H. Dusenov and A. Wibowo, "Deteksi Penipuan Pada Sosial Media Twitter Dengan Metode Bidirectional Long Short Term Memory (Bi-Lstm)," *J. Inform. Manaj. dan Komput.*, vol. 16, no. 1, pp. 117–125, 2024.

[4] Y. Liu, Z. Sun, and W. Zhang, "Improving fraud detection via hierarchical attention-based Graph Neural Network," *J. Inf. Secur. Appl.*, vol. 72, pp. 1–11, 2023, doi: 10.1016/j.jisa.2022.103399.

[5] A. R. Lubis *et al.*, "Deep neural networks approach with transfer learning to detect fake accounts social media on Twitter," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 1, pp. 269–277, 2024, doi: 10.11591/ijeecs.v33.i1.pp269-277.

[6] N. Alharbi, B. Alkalifah, G. Alqarawi, and M. A. Rassam, "Countering Social Media Cybercrime Using Deep Learning: Instagram Fake Accounts Detection," *Futur. Internet*, vol. 16, no. 10, 2024, doi: 10.3390/fi16100367.

[7] M. Ramu, K. Kishorkumar, S. Krishnamoorthi, and C. Praveenbalaji, "International Journal of Research Publication and Reviews AI-Powered Fraud Prevention in Medical Treatment Fundraising," no. 6, pp. 7126–7130, 2025.

[8] Hakim Lukman, "Penerapan Ner Pada Artikel Berita Online Tentang Kriminalitas Menggunakan Indobert Skripsi Program Studi Matematika," 2023.

[9] SANDEEP PAMARTHI, "AI Meets Anonymity: How named entity recognition is redefining data privacy," *World J. Adv. Res. Rev.*, vol. 22, no. 1, pp. 2045–2053, 2024, doi: 10.30574/wjarr.2024.22.1.1270.

[10] B. Acharya, D. Lazzaro, A. E. Cinà, and T. Holz, *Pirates of Charity: Exploring Donation-based Abuses in Social Media Platforms*, vol. 2025, no. v 1.0. Association for Computing Machinery, 2025. doi: 10.1145/3696410.3714634.

[11] S. Imron, E. I. Setiawan, and J. Santoso, "Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN," *J. Intell. Syst. Comput.*, vol. 5, no. 1, pp. 10–16, 2023, doi: 10.52985/insyst.v5i1.267.

[12] X. Yang, S. Saha, A. Venkatesan, S. Tirunagari, V. Vartak, and J. McEntyre, "Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms," *Sci. Data*, vol. 10, no. 1, pp. 1–13, 2023, doi: 10.1038/s41597-023-02617-x.

[13] J. Wang, W. Xu, X. Fu, G. Xu, and Y. Wu, "ASTRAL: Adversarial Trained LSTM-CNN for Named Entity Recognition," *Knowledge-Based Syst.*, vol. 197, 2020, doi: 10.1016/j.knosys.2020.105842.

[14] F. Eko Saputro and A. Fathan Hidayatullah, ©20XX IEEE Tinjauan Literatur : Named Entity Recognition pada Resep Makanan

Indonesia," 2020, [Online]. Available: https://hebbarskitchen.com/
[15] D. S. Rachmad, "Review Named Entity Recognition dengan Menggunakan Machine Learning," *J. Sains dan Inform.*, vol. 6, no. 1, pp. 28–33, 2020, doi: 10.34128/jsi.v6i1.204.
[16] D. Untuk and M. Skripsi, "Named Entity Recognition (Ner) Pada Teks Berbahasa Indonesia Dengan Fine-Tuning Indobert," 2024.
[17] X. Luo, H. Ding, M. Tang, P. Gandhi, Z. Zhang, and Z. He, "Attention Mechanism with BERT for Content Annotation and Categorization of Pregnancy-Related Questions on a Community QA Site," *Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020*, pp. 1077–1081, 2020, doi: 10.1109/BIBM49941.2020.9313379.
[18] S. L. Sariwening, "ABSTRAK IndoBERT: Transformer-based Model for Indonesian Language Understanding," pp. 6–13, 2020, [Online]. Available: http://etd.repository.ugm.ac.id/
[19] F. Zamakhsyari, "Optimasi Model Natural Language Proessing Untuk Aplikasi Question-Answer Kesehatan Mental : Perbandingan Model IndoBERT dan M-BERT," 2024.
[20] N. Istiqomah, F. Novika, S. Tinggi, and M. Asuransi, "Perbandingan Kinerja Model NER IndoBERT dan IndoLEM dalam Ekstraksi Informasi Kesehatan Pascabencana dari Berita Daring di Indonesia," vol. 04, no. 3, pp. 158–174, 2025.
[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
[22] D. Borrelli, G. G. Svartzman, and C. Lipizzi, "Erratum: Unsupervised acquisition of idiomatic units of symbolic natural language: An n-gram frequency-based approach for the chunking of news articles and tweets (PLoS ONE (2020) 15: 6 (e0234214) DOI: 10.1371/journal.pone.0234214)," *PLoS One*, vol. 16, no. 1 January, pp. 1–2, 2021, doi: 10.1371/journal.pone.0245404.