Deteksi Email Phishing Berbasis LSTM yang Ditingkatkan Menggunakan Fitur Emosi yang Diekstraksi oleh DistilBERT

Muhammad Reza Aulia
Fakultas Informatika
Universitas Telkom
Bandung
muauli@students.telkomuniversity.ac.id

Aji Gautama Putrada
Fakultas Informatika
Universitas Telkom
Bandung
Ajigps@telkomuniversity.ac.id

Ikke Dian Oktaviani
Fakultas Informatika
Universitas Telkom
Bandung
oktavianiid@telkomuniversity.ac.id

Abstrak

Apa **: Phishing email merupakan ancaman keamanan siber yang mengeksploitasi aspek psikologis untuk menipu pengguna. Sistem deteksi saat ini umumnya hanya menganalisis pola teks dan mengabaikan fitur manipulasi emosional, sehingga rentan terhadap serangan rekayasa sosial yang canggih. Masukan sistem berupa teks email dan keluarannya adalah klasifikasi phishing atau non-phishing.**

Mengapa **: Topik ini krusial karena serangan phishing sering menjadi pintu masuk serangan siber lainnya. Metode deteksi berbasis deep learning yang ada belum memanfaatkan pemahaman emosi (seperti urgensi atau ketakutan) sebagai fitur deteksi, menciptakan celah keamanan yang signifikan.**

Bagaimana **Penelitian ini mengusulkan metode deteksi yang mengintegrasikan fitur emosional hasil ekstraksi DistilBERT ke dalam klasifikasi Long Short-Term Memory (LSTM). Sistem bekerja dalam dua cabang paralel: satu cabang memproses urutan teks dan cabang lainnya memproses enam label emosi yang diekstrak dari fine-tuned DistilBERT. Kedua fitur tersebut digabungkan untuk menghasilkan prediksi yang lebih akurat.**

Hasil utama**. Evaluasi pada 5.809 email menunjukkan model yang diusulkan mencapai akurasi 0,98 dan recall 0,97, mengungguli model baseline tanpa fitur emosi yang hanya mencapai recall 0,87. Peningkatan recall sebesar 0,10 membuktikan bahwa integrasi analisis emosi efektif mengurangi false negative dalam deteksi phishing.**

Kata kunci: phishing detection, long short-term memory, distilled bidirectional encoder representations from transformers, emotion analysis, deep learning, phishing detection using emotion

Abstract

Email phishing is a significant cyber security threat that exploits emotional manipulation to deceive users. Traditional text-based phishing detection approaches ignore the emotional aspect in their detection systems. This final project proposes an improved phishing detection method by integrating emotional features extracted by DistilBERT with LSTM classification. The dual-branch architecture processes text and emotional features in parallel. DistilBERT, which has been fine-tuned on an emotional dataset, extracts six emotional labels from emails, and these features are then combined with LSTM sequential representations for classification. Evaluation on 5,800 emails showed that the model with emotional features achieved an accuracy of 0.98, while the model without emotional features only achieved an accuracy of 0.96. The biggest improvement was seen in the precision value, which had a difference of 0.03. This explains that the addition of emotional features helps the model to reduce the possibility of mispredictions in detecting emails that are actually safe but are detected as phishing. This is important in a real-world context, as it can reduce user disruption caused by legitimate emails being held as phishing emails. Although the recall value was the same for both models, the improvement in precision and F1-score shows that the model with emotion features provides more balanced and reliable performance. The final project results show that emotion analysis effectively captures patterns of psychological manipulation that are characteristic of social engineering attacks, thereby improving phishing detection performance. The main contribution of this final project is the use of DistilBERT for text emotion detection and its

integration into LSTM-based phishing email detection, which produces optimal performance in reducing prediction errors.

Keywords: phishing detection, long short-term memory, distilled bidirectional encoder representations from transformers, emotion analysis, deep learning, phishing detection using emotion

1. Pendahuluan Latar Belakang

Dalam dunia digital saat ini, organisasi menggunakan berbagai media komunikasi untuk bertukar pesan, gambar, dan dokumen secara cepat [1]. Meskipun demikian, email masih banyak digunakan di lingkungan organisasi karena mendukung komunikasi formal dan pertukaran informasi penting, termasuk data sensitif atau rahasia [2]. Tingginya penggunaan email serta nilai informasi yang dibawanya membuat email menjadi salah satu target utama serangan siber [3]. Salah satu ancaman yang paling umum adalah email phishing, yaitu upaya penipuan yang memanipulasi pengguna agar memberikan informasi sensitif, seperti data keuangan atau kredensial, sehingga dapat menimbulkan kerugian yang signifikan [4]. Selain menimbulkan kerugian langsung, serangan phishing juga kerap menjadi pintu masuk bagi serangan lanjutan, seperti pencurian kredensial dan ransomware [5].

Mengingat adanya celah dalam sistem email, banyak pengguna telah menerapkan filter spam otomatis untuk pertahanan [6]. Namun, efektivitas filter ini terbatas karena email normal sering dikategorikan sebagai spam [7]. Hal ini menjadi penting untuk diperhatikan karena email phishing memiliki karakteristik yang berbeda dari email yang masuk dalam kategori spam [8]. Email phishing dapat menggunakan teks naratif emosional yang menarik dan relevan secara kontekstual dengan pengguna, sehingga menghindari filter spam yang dirancang untuk mendeteksi konten yang berulang atau mencurigakan [9]. Oleh karena itu, pengirim email phishing dapat mengeksploitasi celah ini untuk menurunkan tingkat kewaspadaan pengguna saat membaca email tersebut karena tidak disaring oleh fitur spam [10]. Meskipun tingkat kewaspadaan setiap pengguna tidak selalu rendah, penipu phishing sangat terampil dalam merancang narasi yang memanipulasi korban untuk membuka tautan dalam email melalui manipulasi emosional [11]. Penelitian yang dilakukan oleh Karamagi et al. mengungkapkan bahwa aspek emosional seperti rasa ingin tahu, ketakutan, dan urgensi secara signifikan memengaruhi keputusan korban untuk berinteraksi dengan email berbahaya [12]. Namun, kegagalan untuk mengenali dan menangani manipulasi emosional ini merupakan kelemahan utama dalam sistem pertahanan email saat ini.

Untuk mengatasi masalah ini, peneliti telah menerapkan berbagai pendekatan deep learning untuk deteksi email phishing. Model sekuensial berbasis Recurrent Neural Networks, khususnya Long Short-Term Memory (LSTM) dan variannya, telah menunjukkan keberhasilan signifikan dalam mengklasifikasikan teks email berdasarkan kemampuannya untuk menangkap dependensi temporal jangka panjang dalam urutan teks [13]. Penelitian saat ini telah membuktikan efektivitas arsitektur berbasis LSTM, termasuk Bi-LSTM [14], RNN-LSTM [15], dan hybrid BERT-LSTM [16], yang mampu menganalisis pola sekuensial dan kontekstual dari konten email phishing. Namun, pendekatan berbasis LSTM umumnya masih hanya berfokus pada fitur linguistik dan belum mengeksplorasi fitur manipulasi emosional, yang kini semakin banyak digunakan oleh pelaku phishing.

Di sisi lain, arsitektur deep learning berbasis Transformer telah merevolusi pemrosesan bahasa alami, menawarkan kemampuan pemahaman kontekstual yang superior melalui mekanisme self-attention [17]. Model Transformer telah menunjukkan keberhasilan signifikan dalam berbagai tugas Natural Language Processing (NLP), termasuk klasifikasi emosi dalam data teks, dengan kemampuan menangkap hubungan antara setiap kata dan konteks emosional yang kompleks. Kemampuan ini memberikan peluang baru untuk meningkatkan deteksi phishing dengan mengintegrasikan pemahaman emosi sebagai dimensi tambahan dalam analisis. Namun, penerapan model Transformer dalam deteksi phishing sering berfokus pada klasifikasi biner seperti phishing atau non-phishing secara langsung tanpa memanfaatkan kemampuan pemahaman emosi sebagai fitur tambahan [18]. Dengan demikian, terdapat gap penelitian yang signifikan untuk mengintegrasikan analisis emosi yang diekstraksi oleh model Transformer ke dalam alur kerja deteksi phishing berbasis LSTM, sehingga meningkatkan kemampuan

mengidentifikasi serangan yang memanfaatkan manipulasi psikologis. Berdasarkan permasalahan di atas, tugas akhir ini mengusulkan metode deteksi phishing yang menggunakan emosi sebagai pertimbangan untuk deteksi phishing.

Tahap pertama berfokus pada klasifikasi teks email phishing menggunakan model Transformer, yaitu DistilBERT, model Transformer yang lebih ringan dan cepat yang tetap mempertahankan performa tinggi dari BERT [19]. Tugas akhir ini melakukan fine-tune DistilBERT untuk tugas klasifikasi emosi, dan kemudian menggunakannya untuk pelabelan dalam dataset email phishing berdasarkan emosi yang telah dilakukan fine tune dari dataset teks emosi untuk setiap email dalam dataset. Pada tahap kedua, dataset ini diintegrasikan sebagai input tambahan ke dalam model LSTM, yang dilatih untuk tugas klasifikasi phishing akhir. Untuk memvalidasi efektivitas pendekatan ini, dilakukan perbandingan kinerja model LSTM yang menggunakan label emosi dengan model LSTM yang hanya menggunakan data teks. Selama melakukan riset, saya belum menemukan penelitian yang menggunakan deteksi emosi melalui pemrosesan teks untuk meningkatkan performa deteksi email phishing. Kontribusi utama dari tugas akhir ini adalah penggunaan DistilBERT untuk deteksi emosi teks dan menyisipkannya dalam deteksi email phishing berbasis LSTM.

Topik dan Batasannya

Penelitian ini membandingkan teknik klasifikasi email phishing melalui integrasi fitur emosi hasil prediksi DistilBERT sebagai fitur tambahan pada model LSTM dengan model tanpa fitur emosi pada model LSTM. Sistem menerima input berupa konten teks email dan menghasilkan output klasifikasi biner (email normal atau phishing). Sebagai contoh, email dengan teks "Klik link berikut untuk verifikasi akun Anda" diproses untuk mengekstrak fitur teks melalui DistilBERT, dilanjutkan dengan prediksi emosi dari enam kategori (Anger, Joy, Love, Sadness, Fear, Surprise), kemudian kedua fitur tersebut digabungkan sebagai input LSTM untuk menghasilkan prediksi klasifikasi final. Penelitian mengeksplorasi kontribusi fitur emosi dalam meningkatkan akurasi deteksi phishing dengan membandingkan performa model LSTM dalam dua skenario: tanpa fitur emosi dan dengan integrasi fitur emosi, menggunakan metrik evaluasi Accuracy, Precision, Recall, dan F1-Score.

Batasan masalah ditetapkan untuk memastikan agar cakupan penelitian ini jelas. Dataset terbatas pada 5.800 sampel email publik dari Kaggle (4.091 normal, 1.709 phishing) karena keterbatasan kapasitas komputasi yang tersedia. Klasifikasi emosi dibatasi enam kategori karena ketersediaan dataset emosi untuk fine-tuning DistilBERT dan keterbatasan komputasi. Evaluasi hanya pada konten teks email tanpa metadata lain (header, IP, lampiran) karena ekstraksi metadata memerlukan infrastruktur tambahan tidak tersedia dalam jangka waktu semester. Model DistilBERT menggunakan versi base-uncased tanpa modifikasi arsitektur lebih lanjut karena keterbatasan waktu dan resources. Evaluasi performa menggunakan metrik standar klasifikasi biner pada dataset pengujian terpisah karena terstandar dan dapat dibandingkan dengan penelitian sejenis. Batasan-batasan ini dapat dilonggarkan pada penelitian lanjutan dengan meningkatkan kapasitas komputasi, memperluas jumlah dan keragaman dataset, mengintegrasikan metadata email, atau mengoptimalkan arsitektur model.Tujuan

Tujuan penelitian ini adalah menganalisis kontribusi fitur emosi terhadap performa deteksi email phishing melalui pendekatan integrasi fitur emosi hasil prediksi DistilBERT pada arsitektur model LSTM. Penelitian membandingkan performa model LSTM dalam dua skenario eksperimen: skenario pertama menggunakan fitur tekstual saja, dan skenario kedua mengintegrasikan fitur emosi bersama fitur tekstual. Kedua skenario dijalankan dengan dataset dan pengaturan pelatihan yang identik untuk memastikan validitas perbandingan. Analisis performa dilakukan menggunakan metrik evaluasi Accuracy, Precision, Recall, dan F1-Score pada dataset pengujian terpisah untuk mengukur pengaruh nyata penambahan fitur emosi terhadap kemampuan model mengenali email phishing. Hasil perbandingan ini akan menunjukkan apakah fitur emosi secara signifikan meningkatkan performa deteksi atau tidak.

Organisasi Tulisan

Sistematika penulisan jurnal ini disusun sebagai berikut: Bagian 2 membahas studi terkait yang menjadi landasan teori dan penelitian terdahulu mengenai sistem deteksi phising tradisional hingga deteksi phishing menggunakan deep learning dan transformer. Bagian 3 menjelaskan metodologi penelitian dan sistem yang dibangun, mencakup deskripsi dataset Teks emosi dan email phishing, tahapan pre-processing data, serta konfigurasi fine-tuning model DistilBERT dan konfigurasi model LSTM. Bagian 4 memaparkan hasil eksperimen yang meliputi evaluasi berdasarkan angka yang dihasilkan dari metrik evaluasi dan analisis mengapa model emosi lebih baik dibandingkan model tanpa fitur emosi. Terakhir, Bagian 5 menyajikan kesimpulan yang diperoleh dari penelitian beserta saran untuk pengembangan di kedepannya.

2. Studi Terkait

Deteksi email phishing menggunakan pendekatan deep learning telah menjadi fokus penelitian dengan berbagai metode dan arsitektur model. Penelitian terdahulu menunjukkan bahwa kombinasi model dan penambahan fitur memberikan peningkatan performa deteksi. Rokunojjaman et al. menggunakan pendekatan hybrid dengan menggabungkan CNN dan LSTM untuk ekstraksi fitur dan klasifikasi phishing email, mencapai akurasi 0.96 [20]. Rafat et al. membandingkan model deep learning dengan machine learning tradisional, menemukan bahwa LSTM mencapai akurasi tertinggi 0.90 untuk deteksi phishing email [21]. Asani et al. mengembangkan sistem deteksi phishing berbasis API menggunakan kombinasi LSTM dan MLP dengan akurasi 0.97 [22]. Bhatti et al. menerapkan teknik penyeimbangan kelas pada model berbasis LSTM untuk mengatasi ketidakseimbangan data, mencapai akurasi 0.95 [23].

TABEL 1
Studi Terkait tentang Deteksi email phishing

Reference	Model/Pendekatan	Contribution	*Accuracy*
Rokunojjaman *et al.* [20]	LSTM–CNN	Menggabungkan CNN dan LSTM (hybrid) untuk deteksi phishing *email*	0.96
Rafat *et al.* [21]	LSTM dan	Membandingkan model *deep learning* vs machine learning tradisional untuk deteksi phishing *email*, dengan model LSTM memiliki akurasi tertinggi	0.90
Asani *et al.* [22]	LSTM + MLP,berbasis API	Menyediakan phishing detection berbasis API ,memakai LSTM dan MLP	0.97
Bhatti *et al.* [23]	LSTM + balancing	Menambahkan teknik penyeimbangan (balancing) pada deteksi berbasis LSTM	0.95

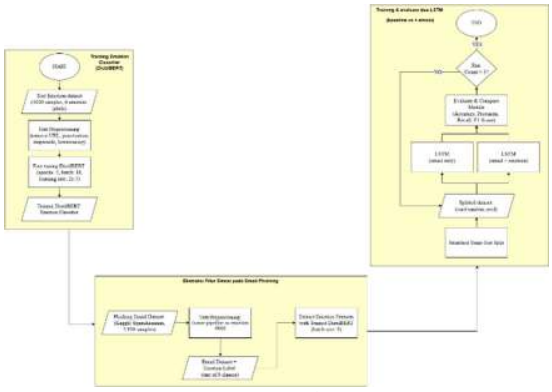
Tren umum dalam penelitian sebelumnya adalah peningkatan performa melalui penggabungan model (hybrid architecture), penambahan teknik preprocessing atau balancing, serta implementasi berbasis infrastruktur seperti API. Namun, sebagian besar pendekatan masih terfokus pada fitur tekstual dan teknis email, sementara aspek emosi yang sering dimanfaatkan pelaku phishing untuk memengaruhi korban belum diintegrasikan secara eksplisit sebagai fitur dalam proses klasifikasi. Penelitian ini mengisi gap tersebut dengan mengintegrasikan fitur emosi hasil prediksi DistilBERT sebagai fitur tambahan pada model LSTM, sehingga model dapat menangkap pola perilaku linguistik dan emosional yang kompleks dalam email phishing. Metrik evaluasi yang

digunakan mengacu pada standar klasifikasi biner: Accuracy, Precision, Recall, dan F1-Score .

3. Sistem Yang Dibangun

3.1 Desain Perancangan Sistem

Arsitektur yang diusulkan beroperasi dalam tiga tahap yang berbeda, seperti yang ditunjukkan pada Gambar 2. Tahap pertama melakukan penyempurnaan model DistilBERT dengan cara di *fine tune* pada dataset teks emosional untuk menghasilkan klasifikasi emosi yang terlatih. Model DistilBERT dipilih karena efisiensinya yang tinggi—ukuran model ini sekitar 40% lebih kecil daripada BERT sambil tetap mempertahankan 97% kemampuan pemahaman bahasa—sehingga cocok untuk implementasi dengan sumber daya terbatas [19]. Tahap kedua menggunakan model DistilBERT yang telah di *fine tune* untuk mengekstrak fitur emosional dari setiap email dalam dataset phishing. Proses ekstraksi dilakukan secara batch dengan ukuran batch 8 untuk efisiensi memori, menghasilkan dataset yang diperkaya dengan anotasi emosi dari enam kategori: Marah, Kebahagiaan, Cinta, Kesedihan, Ketakutan, dan Keterkejutan. Ekstraksi ini menghasilkan 5.800 email yang dilabeli dengan emosi, membentuk dataset phishing yang diperluas untuk tahap klasifikasi. Selanjutnya tahap ketiga terjadi proses pelatihan dan evaluasi pada dua model LSTM. Model pertama adalah model dasar, atau model deteksi, yang tidak menggunakan fitur emosional untuk deteksi dan hanya menggunakan fitur teks. Model kedua adalah model yang ditingkatkan dengan fitur emosional, yang menggabungkan fitur teks dan emosional untuk mengukur kontribusi informasi emosional terhadap kinerja deteksi phishing. Kedua model dilatih dengan hiperparameter yang identik dan dataset yang sama untuk memastikan perbandingan yang adil, sehingga peningkatan kinerja apa pun dapat langsung dikaitkan dengan penambahan fitur emosional.



GAMBAR 1
Flowchart arsitektur deteksi phishing dan fitur ekstraksi

3.2 Dataset dan *Pre-processing*

Dalam penelitian ini, kami menggunakan dataset teks emosional yang tersedia di platform Kaggle, yaitu ‘nelgiriyewithana/emotions’.[24]. Dataset ini berisi sekitar 40.000 sampel teks emosional yang dilabeli sebagai ‘kemarahan’, ‘kegembiraan’, ‘bahagia’, ‘kesedihan’, ‘ketakutan’, dan ‘kejutan’. Untuk dataset email phishing, kami mengambilnya dari platform Kaggle, yaitu ‘naserabdullahalam/phishing-email-dataset’ [25],yang secara khusus menggunakan file dataset email phishing. Dataset ini berisi 5.809 contoh email dengan distribusi 4.091 email biasa dan 1.718 email phishing. Dataset ini mencakup metadata email yang komprehensif, termasuk pengirim, penerima, cap waktu, subjek, isi pesan, label klasifikasi, dan informasi URL, yang memberikan konteks lengkap untuk analisis phishing. Dataset tersebut kemudian diekstraksi fiturnya menggunakan DistilBERT, yang telah disesuaikan untuk deteksi emosi. Hasilnya adalah dataset yang telah diintegrasikan dengan fitur emosi untuk deteksi oleh model LSTM. Proses *pre-processing* pada dataset menerapkan transformasi konsisten pada dataset teks emosional dan dataset email phishing untuk memastikan representasi teks yang homogen dan meminimalkan noise. Pipelinenya terdiri dari enam langkah berurutan, yaitu penghapusan URL menggunakan pencocokan pola ekspresi reguler untuk menghilangkan tautan Hypertext Transfer Protocol/Hypertext Transfer Protocol

Secure (HTTP/HTTPS) dan alamat web, diikuti dengan penghapusan karakter khusus dan tanda baca menggunakan penyaringan karakter berbasis ekspresi reguler (regex), kemudian normalisasi spasi dengan mengurangi spasi berurutan menjadi satu spasi, penghapusan nilai numerik untuk menghilangkan noise dari angka, konversi ke huruf kecil untuk menstandarkan representasi teks, dan akhirnya penghapusan kata-kata stop bahasa Inggris menggunakan korpus kata-kata stop bahasa Inggris dari Natural Language Toolkit (NLTK). Pendekatan prapemrosesan ini menghapus informasi yang tidak relevan (URL, karakter khusus, angka, kata stop) dan hanya mempertahankan kata-kata yang memiliki nilai semantik.

Baris dengan teks kosong setelah prapemrosesan dihapus untuk memastikan integritas dataset. Yang paling penting, pipeline prapemrosesan ini diterapkan secara identik pada kedua dataset yaitu dataset emosi untuk fine-tuning DistilBERT, dan dataset email phishing untuk deteksi dan penandaan fitur emosi. Konsistensi ini memastikan bahwa model DistilBERT dapat secara akurat mengekstrak fitur emosi dari email phishing, karena format teks email sama dengan format teks emosi yang digunakan selama pelatihan.

3.3 Konfigurasi Fine-Tunning DistilBERT untuk klasifikasi emosi

DistilBERT-base-uncased digunakan sebagai model dasar untuk klasifikasi emosi karena memiliki efisiensi komputasi yang tinggi dengan kemampuan untuk mempertahankan kinerja yang kuat dan andal dalam memproses data emosi. Konfigurasi model DistilBERT mencakup 6 label output yang sesuai dengan kategori emosi dalam dataset emosi untuk pelatihan. Selain itu, digunakan tokeniser dengan padding maksimum 128 token, diikuti dengan pemotongan untuk menstandarkan panjang urutan input di seluruh dataset pelatihan. Proses pelatihan model dilakukan menggunakan Trainer API dari perpustakaan Transformers. Model dioptimalkan dengan laju pembelajaran 2×10^-5, ukuran batch 16 pada fase pelatihan dan evaluasi, serta total 3 epoch pelatihan. Parameter weight decay ditetapkan sebesar 0.01 untuk mengurangi overfitting, dan proses evaluasi dilakukan pada akhir setiap epoch. Kinerja model dievaluasi menggunakan metrik akurasi, F1-score, presisi, dan recall, di mana semua metrik dihitung menggunakan pendekatan rata-rata yang seimbang untuk memperhitungkan ketidakseimbangan distribusi kelas dalam data pelatihan. Setelah proses penyempurnaan selesai, model DistilBERT yang telah dilatih akan mengekstrak fitur emosional dari dataset email phishing yang telah melalui prapemrosesan yang dioptimalkan dengan ukuran batch 8 untuk efisiensi memori. Setiap email akan menjalani prapemrosesan yang sama seperti pada pelatihan sebelumnya, diikuti dengan prediksi emosi menggunakan model yang telah disempurnakan sebelumnya. Hasil dari proses penandaan ini adalah dataset yang berisi konten email phishing yang dilengkapi dengan anotasi fitur emosional yang sesuai.

Selanjutnya Dataset teks emosional dan email phishing diproses melalui tokenisasi menggunakan tokenizer dengan Vocabulary size sebesar 5000 dan penanganan kata di luar kamus untuk menghilangkan kata-kata pendek, kesalahan ketik, dan sebagainya. Penambahan urutan (sequence padding) diterapkan dengan panjang maksimum yang berbeda untuk setiap tahap, yaitu 128 token untuk pelatihan DistilBERT yang mempelajari dataset teks emosi, yang ditentukan berdasarkan analisis distribusi panjang teks dalam masing-masing data set. Fitur emosi diproses menggunakan OneHotEncoder, yang mengubah enam kategori emosi, yaitu Kesedihan, Kebahagiaan, Cinta, Kemarahan, Ketakutan, dan Keterkejutan, menjadi vektor fitur biner enam dimensi. Misalnya, emosi ‘Kebahagiaan’ diwakili sebagai ([0, 1, 0, 0, 0, 0]), di mana setiap posisi mewakili satu kategori emosi. Proses enkoding ini menghasilkan representasi numerik yang kompatibel dengan arsitektur jaringan saraf dan memungkinkan model untuk memproses informasi teks dan emosi secara bersamaan.

3.4 Arsitektur Long Short-Term Memory (LSTM)

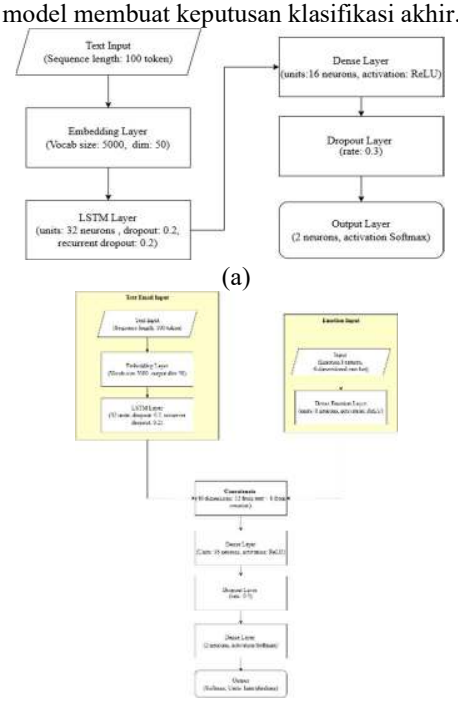
3.4.1 LSTM Baseline Model

Model LSTM dasar hanya memproses

fitur teks tanpa memasukkan fitur emosional, kemudian menerapkan arsitektur satu cabang seperti yang ditunjukkan pada Gambar 3(a). Arsitektur ini dimulai dengan lapisan Input yang menerima urutan teks berukuran 100 token, diikuti oleh lapisan Embedding dengan ukuran kamus 5.000 dan dimensi keluaran 50 yang mengubah token menjadi representasi vektor padat. Lapisan LSTM dengan 32 unit, tingkat dropout 0,2, dan dropout rekurens 0,2 menangkap ketergantungan urutan dalam teks email. Keluaran LSTM kemudian diproses melalui lapisan Dense pertama dengan 16 unit dan aktivasi ReLU untuk transformasi fitur non-linier, diikuti oleh lapisan Dropout dengan tingkat 0,3 untuk regularisasi. Lapisan output Dense dengan 2 unit dan aktivasi Softmax menghasilkan probabilitas klasifikasi biner untuk kelas ham dan phishing. Konfigurasi ini menetapkan konfigurasi awal untuk kinerja yang dapat dicapai dengan fitur teks saja, dan berfungsi sebagai perbandingan dengan model yang ditingkatkan dengan emosi. Model dilatih menggunakan optimiser Adam dengan laju pembelajaran 0.001 dan fungsi kerugian cross-entropy kategorikal, dengan penghentian dini untuk mencegah overfitting.

3.4.1 LSTM Emotion Model

Model yang diperkaya emosi menerapkan arsitektur dua cabang yang memproses informasi teks dan emosi secara paralel, seperti yang ditunjukkan pada Gambar 3(b). Arsitektur ini terdiri dari dua cabang input yang berbeda, yaitu cabang input teks dan cabang input emosi. Cabang teks dalam model ini mengikuti struktur yang sama dengan model dasar, yaitu 100 token input teks diproses melalui lapisan embedding dengan 5.000 kata dalam kamus, dengan dimensi output 50. Ini diikuti oleh lapisan LSTM dengan 32 unit, dropout 0,2, dan recurrent dropout 0,2, yang bertugas menangkap ketergantungan teks secara berurutan. Output dari cabang teks ini menghasilkan representasi vektor berdimensi 32. Cabang emosi memproses vektor emosi yang dienkode menggunakan enkode one-hot enam dimensi secara terpisah. Vektor-vektor ini mewakili enam kategori emosi yang diekstraksi oleh model DistilBERT (Kemarahan, Kebahagiaan, Cinta, Kesedihan, Ketakutan, Keterkejutan). Cabang emosi menggunakan lapisan Dense dengan 8 unit dan aktivasi ReLU untuk mengekstraksi representasi emosi yang padat dari masukan enkode one- hot. Selama tahap integrasi dan klasifikasi, keluaran dari kedua cabang digabungkan menggunakan operasi concatenate. Cabang teks menghasilkan representasi berdimensi 32, sementara cabang emosi menghasilkan representasi berdimensi 8, sehingga penggabungan keduanya membentuk representasi terpadu berdimensi 40. Representasi gabungan ini kemudian diproses melalui serangkaian lapisan klasifikasi dengan struktur yang identik dengan model dasar. Lapisan Dense pertama dengan 16 unit dan aktivasi Rectified Linear Unit (ReLU) melakukan transformasi fitur non-linier, diikuti oleh lapisan dropout dengan tingkat 0,3 untuk regularisasi, dan akhirnya lapisan output dense dengan 2 unit dan aktivasi Softmax menghasilkan probabilitas klasifikasi biner. Arsitektur dua cabang ini memungkinkan model untuk memanfaatkan informasi dari dua jenis data input yang berbeda. Polanya teks tingkat rendah yang ditangkap oleh cabang LSTM diintegrasikan dengan indikator semantik tingkat tinggi yang dienkode dalam cabang emosi sebelum model membuat keputusan klasifikasi akhir.



(a)

(b)

GAMBAR 2
Phishing detection models: (a) baseline model and (b) emotion enhanced model

3.5 Optimasi Model LSTM dan Strategi Pelatihan

Kedua model dilatih menggunakan perhitungan hiperparameter yang sama agar dapat dibandingkan secara adil. Perhitungan hiperparameter menggunakan 25 epoch, dengan ukuran batch 64 untuk setiap batch. Proses pelatihan menggunakan optimiser Adam dengan laju pembelajaran 0.001, yang berfungsi sebagai pengatur laju pembelajaran untuk setiap parameter secara adaptif berdasarkan gradien yang dihitung, sehingga mempercepat konvergensi dan meningkatkan stabilitas selama proses pelatihan. Fungsi kerugian yang digunakan adalah categorical cross-entropy, yang merupakan metode untuk mengukur kesalahan prediksi dalam tugas klasifikasi antara email phishing dan non-phishing. Fungsi ini menghitung tingkat kesalahan prediksi yang dihasilkan oleh model berdasarkan label dalam dataset dan menghukum model ketika model sangat yakin dengan prediksinya tetapi ternyata salah dibandingkan dengan label dalam dataset. Selama proses pelatihan, kumpulan data validasi menggunakan kumpulan data uji yang dibagi. Kumpulan data akan dibagi menjadi kumpulan data pelatihan dan kumpulan data uji dengan perbandingan 80 persen untuk pelatihan dan 20 persen untuk pengujian. Pembagian ini menggunakan sampling terstratifikasi untuk mempertahankan distribusi kelas yang sama. Stratifikasi sangat penting dalam kasus kumpulan data phishing yang memiliki ketidakseimbangan kelas, seperti jumlah email phishing yang lebih sedikit dibandingkan dengan email non-phishing. Mekanisme penimbangan kelas juga diterapkan untuk memberikan penalti yang lebih besar saat mengklasifikasikan kelas yang jumlahnya sedikit dalam dataset ini, artinya ketika model mendeteksi email phishing, model akan lebih sensitif daripada saat mendeteksi email ham. Hal ini dapat terjadi karena mekanisme penimbangan kelas memberikan "berat" yang lebih besar pada model saat mendeteksi email phishing. Selama proses pelatihan, setiap epoch akan memproses data pelatihan yang telah dibagi sebelumnya dalam pemrosesan batch, lalu menghitung jumlah kerugian yang terjadi pada setiap epoch. Selanjutnya, model akan belajar memperbaiki kesalahannya sendiri untuk meningkatkan akurasi berdasarkan penalti kesalahan yang lebih tinggi untuk kesalahan prediksi pada email phishing, yang merupakan kelas minoritas. Pada akhir setiap siklus epoch, hasil akan diuji berdasarkan data validasi, dan mekanisme penghentian dini juga akan memeriksa apakah terdapat peningkatan akurasi selama 7 epoch berturut-turut. Jika tidak ada peningkatan akurasi selama 7 epoch berturut-turut, mekanisme penghentian dini akan diaktifkan dan menghentikan proses untuk mencegah overfitting, serta menggunakan hasil dengan akurasi validasi tertinggi yang telah disimpan selama proses pelatihan untuk digunakan dalam metrik evaluasi.

4. Hasil penelitian dan perbandingan model

4.1 Hasil Pengujian

Percobaan dilakukan pada dataset *email* phishing berjumlah 5.800 sampel yang telah melalui prapemrosesan. Sebanyak 9 baris data dihapus karena memiliki nilai kosong. Evaluasi dilakukan dengan membandingkan dua model, yaitu model baseline (tanpa fitur emosi) dn model dengan fitur emosi. Untuk memperoleh hasil yang lebih stabil, pengujian dijalankan sebanyak 3 kali dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian pada setiap run, serta menggunakan random state yang berbeda. Nilai pada Tabel 4.1 merupakan rata-rata dari ketiga run. Metrik evaluasi yang digunakan meliputi *accuracy*, precision, recall, dan F1- score.

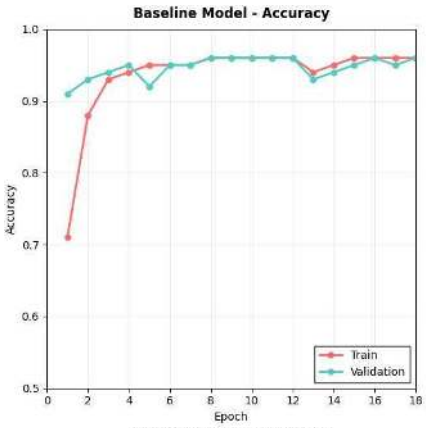
TABEL 2
Perbandingan model baseline dengan model yang ditingkatkan dengan emosi

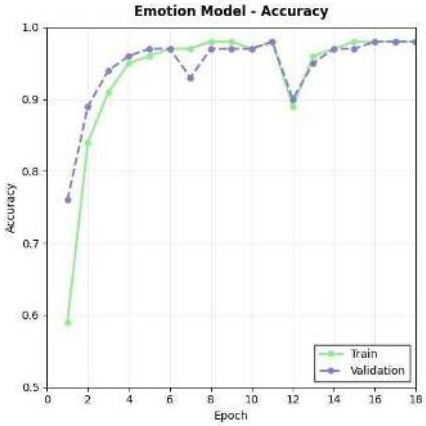
Metrics	Performa Model	
	Dengan Fitur Emosi	Tanpa Fitur Emosi

	(Model 1)	(Model 2)
Accuracy	0.98	0.96
Precision	0.97	0.94
Recall	0.95	0.95
F1-score	0.96	0.94

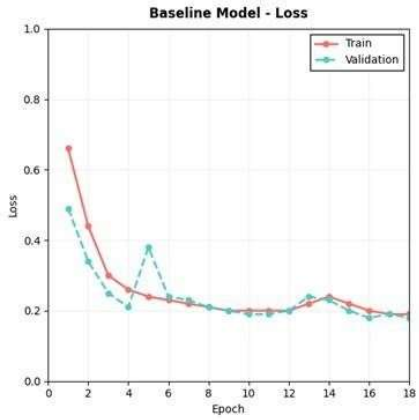
4.2 Analisis Hasil Pengujian

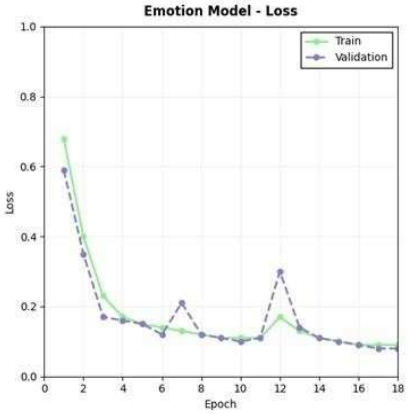
Berdasarkan rata-rata dari 3 kali pengujian, model dengan fitur emosi menunjukkan performa yang lebih baik dibandingkan model baseline. Model dengan fitur emosi mencapai akurasi 0,98, sedangkan baseline 0,96, yang menunjukkan peningkatan ketepatan prediksi dalam mengklasifikasikan *email* sebagai phishing atau non-phishing. Peningkatan paling menonjol terlihat pada precision, yaitu 0,97 pada model dengan fitur emosi dan 0,94 pada baseline. Hasil ini menunjukkan bahwa penambahan fitur emosi membuat model lebih tepat ketika memberikan label phishing, sehingga mengurangi kesalahan saat *email* yang sebenarnya aman justru ditandai sebagai *email phishing*. Jika model ini diterapkan di kehidupan nyata, tingkat precision yang tinggi ini penting karena dapat mengurangi gangguan pada pengguna, misalnya tertahannya *email* yang harusnya bukan *phishing* namun di deteksi sebagai *email phishing*. Selanjutnya adalah meningkatnya nilai F1-score dari 0,94 pada model baseline menjadi 0,96 pada model dengan fitur emosi. Karena F1-score merepresentasikan keseimbangan antara precision dan recall, peningkatan ini menunjukkan performa yang lebih seimbang, terutama didorong oleh kenaikan precision. Berdasarkan tabel diatas, nilai recall pada kedua model tercatat sama, yaitu 0.95. Artinya, proporsi *email* phishing yang berhasil terdeteksi tidak meningkat maupun menurun setelah penambahan fitur emosi. Dapat dikatakan bahwa kontribusi utama fitur emosi pada percobaan ini lebih terlihat pada peningkatan ketepatan prediksi phishing daripada meningkatkan jumlah phishing yang berhasil di prediksi.





(a)





(b)

GAMBAR 3
Training dynamics: (a) kurva akurasi dan (b) kurva loss

Hasil percobaan pada Tabel 2 diperkuat oleh kurva pelatihan pada Gambar 4. Pada kurva loss, kedua model menunjukkan penurunan loss yang cepat pada epoch awal dan kemudian stabil, yang mengindikasikan proses pembelajaran berjalan dengan baik. Model dengan fitur emosi cenderung mencapai kondisi stabil lebih awal sehingga lebih cepat mencapai performa optimal. Pada kurva akurasi, kedua model sama-sama mencapai akurasi yang tinggi; namun, model dengan fitur emosi lebih konsisten mempertahankan akurasi pada data training maupun validasi. Secara keseluruhan, penambahan fitur emosi memberikan dampak positif terutama pada peningkatan precision dan F1-score. Sementara itu, nilai recall yang sama pada kedua model menunjukkan bahwa kontribusi fitur emosi lebih terlihat pada ketepatan prediksi phishing, bukan pada peningkatan jumlah phishing yang terdeteksi.

5. Kesimpulan

Penelitian ini mengusulkan model deteksi phishing yang diperkaya dengan emosi, yang mengintegrasikan fitur emosional yang diekstraksi oleh DistilBERT dan diklasifikasikan oleh model LSTM sebagai model deteksi phishing. Model ini dievaluasi dan dilatih menggunakan dataset yang terdiri dari 5.800 sampel email untuk melakukan klasifikasi biner antara email phishing dan non-phishing. Penerapan fitur emosional ke dalam model deteksi phishing meningkatkan akurasi sebesar 0,03 dari 0,95 tanpa fitur emosional menjadi 0,98 saat diintegrasikan dengan fitur emosional. Peningkatan paling signifikan terlihat pada skor recall, yang meningkat sebesar 0,1 dari 0,87 sebelum integrasi menjadi 0,97 setelah diintegrasikan dengan fitur emosional. Peningkatan 0,1 pada recall menunjukkan bahwa model yang diintegrasikan dengan fitur emosional berhasil mengurangi false negatives. Model dengan fitur emosional tambahan juga menunjukkan konvergensi yang lebih cepat dan stabil dibandingkan dengan model dasar selama pelatihan model. Hal ini membuktikan bahwa analisis emosional dapat membantu meningkatkan deteksi phishing, karena manipulasi emosional juga merupakan karakteristik kunci dari serangan rekayasa sosial. Penelitian ini berkontribusi pada deteksi phishing dengan menunjukkan bahwa menggabungkan dua masukan, yaitu teks email dan fitur emosional, menghasilkan kinerja yang lebih baik daripada pendekatan dengan satu masukan. Model yang diusulkan memberikan solusi yang lebih komprehensif untuk mendeteksi serangan rekayasa sosial yang memanipulasi emosi. Namun, masih ada beberapa keterbatasan, termasuk ukuran dataset yang terbatas dan potensi untuk meningkatkan deteksi emosi dengan menggunakan arsitektur yang lebih canggih. Penelitian di masa depan diharapkan untuk mengeksplorasi dataset yang lebih besar dan beragam yang mencakup tren phishing terbaru. Selain itu, mengeksplorasi model transformer atau model lain yang lebih baik diharapkan dapat meningkatkan akurasi ekstraksi fitur emosional.

REFERENSI

[1] A. Dhakal, “The Need for a Multi-Media Approach in Agricultural Communication: Bridging Socioeconomic Gaps,” *European Journal of Theoretical and Applied Sciences*, vol. 3, no. 1, pp. 3–10, Jan. 2025, doi: 10.59324/ejtas.2025.3(1).01.

[2] A. Jean and T. Alherbe, “Gid Crypto: Application for End-to-End Encrypt and Decrypt E-mail and Data,” *ASEAN Journal of Scientific and Technological Reports*, vol. 27, no. 2, pp. 90–102, Mar.

2024, doi: 10.55164/ajstr.v27i2.251127.

[3] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, Mar. 2021, doi: 10.3389/fcomp.2021.563060.

[4] N. PETLIAK, Y. BEZKOROVALNYI, and N. KUPCHYK, "ANALYSIS OF MODERN METHODS OF DETECTION OF PHISHING E-MAILS," *Herald of Khmelnytskyi National University. Technical sciences*, vol. 341, no. 5, pp. 510–515, Oct. 2024, doi: 10.31891/2307-5732-2024-341-5-73.

[5] M. Benmalek, "Ransomware on cyber-physical systems: Taxonomies, case studies, security gaps, and open challenges," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 186–202, 2024, doi: 10.1016/j.iotcps.2023.12.001.

[6] S. Eppakayala and S. K. G. Ankula, "Intelligent Spam Filtering and Analysis Using Web Automation, ReportLab, and Ensemble Machine Learning Techniques," Feb. 11, 2025. doi: 10.21203/rs.3.rs-5941111/v1.

[7] T. Amutha and S. Geetha, "Automated Spam Detection Using Sandpiper Optimization Algorithm-Based Feature Selection with the Machine Learning Model," *IETE J. Res.*, vol. 70, no. 2, pp. 1472–1479, Feb. 2024, doi: 10.1080/03772063.2023.2280663.

[8] G. R and F. M. Philip, "Phishing Detection," *YMER Digital*, vol. 21, no. 06, pp. 405–412, Jun. 2022, doi: 10.37896/YMER21.06/39.

[9] A. G. Putrada and R. F. Aushaf, "L2 Regularization to Mitigate Overfitting in Email Phishing Attack Detection by CNN-LSTM," in *2024 IEEE 22nd Student Conference on Research and Development (SCOReD)*, IEEE, Dec. 2024, pp. 416–421. doi: 10.1109/SCOReD64708.2024.10872755.

[10] M. Jari, "A Comprehensive Survey of Phishing Attacks and Defences: Human Factors, Training and the Role of Emotions," *International Journal of Network Security & Its Applications*, vol. 14, no. 5, pp. 11–24, Sep. 2022, doi: 10.5121/ijnsa.2022.14502.

[11] Santosa Pohan, Desi Irfan, Intan Nur Fitriyani, Yusril Iza Mahendra Hasibuan, and Indah Chayani, "Simulation and Detection of Phishing Attacks on Student Academic Emails Using Social Engineering Techniques," *International Journal of Health Engineering and Technology*, vol. 2, no. 4, May 2025, doi: 10.55227/ijhet.v2i4.283.

[12] R. Karamagi and S. Ally, "Security Risk Scale: A Case of Email Phishing Detection using Text Mining," *Journal of ICT Systems*, vol. 1, no. 2, pp. 93–108, Dec. 2023, doi: 10.56279/jicts.v1i2.43.

[13] A. J. J S, "A Survey on Phishing Email Detection Techniques: Using LSTM and Deep Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 8, pp. 1597–1600, Aug. 2025, doi: 10.22214/ijraset.2025.73836.

[14] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," *Sensors*, vol. 24, no. 7, p. 2077, Mar. 2024, doi: 10.3390/s24072077.

[15] O. Senouci and N. Benaouda, "Enhancing Phishing Detection in Cloud Environments Using RNN-LSTM in a Deep Learning Framework," *Journal of Telecommunications and Information Technology*, no. 1, pp. 1–9, Mar. 2025, doi: 10.26636/jtit.2025.1.1916.

[16] S. Atawneh and H. Aljehani, "Phishing Email Detection Model Using Deep Learning," *Electronics (Basel).*, vol. 12, no. 20, p. 4261, Oct. 2023, doi: 10.3390/electronics12204261.

[17] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5789–5829, Dec. 2021, doi: 10.1007/s10462-021-09958-2.

[18] S. Jamal, H. Wimmer, and I. H. Sarker, "An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach," *SECURITY AND PRIVACY*, vol. 7, no. 5, Sep. 2024, doi: 10.1002/spy2.402.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of

BERT: smaller, faster, cheaper and lighter," Mar. 2020.

[20] M. Rokunojjaman, A. Malik, M. M. J. Sajeeb, M. Yahiduzzaman, and M. S. Islam, "Multi-Feature Hybrid LSTM-CNN Framework for Phishing Email Detection," *International Journal of Electronics and Communications Systems*, vol. 5, no. 1, pp. 93–103, Jun. 2025, doi: 10.24042/ijecs.v5i1.26764.

[21] K. F. Rafat, Q. Xin, A. R. Javed, Z. Jalil, and R. Z. Ahmad, "Evading obscure communication from spam emails," *Mathematical Biosciences and Engineering*, vol. 19, no. 2, pp. 1926–1943, 2021, doi: 10.3934/mbe.2022091.

[22] E. O. Asani, O. S. Babalola, A. E. Akinola, A. S. Barnabas, O. V. Adams, and J. O. Odumesi, "A Server-side Phishing detection API using Long-Short Term Memory (LSTM) and Multi-layer Perceptron (MLP)," in *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, IEEE, Apr. 2024, pp. 1–8. doi: 10.1109/SEB4SDG60871.2024.10629792.

[23] P. Bhatti, Z. Jalil, and A. Majeed, "Email Classification using LSTM: A Deep Learning Technique," in *2021 International Conference on Cyber Warfare and Security (ICCWS)*, IEEE, Nov. 2021, pp. 100–105. doi: 10.1109/ICCWS53234.2021.9703084.

[24] N. Elgiriyewithana, "Emotions dataset for NLP," Kaggle. Accessed: Nov. 13, 2025. [Online]. Available: https://www.kaggle.com/datasets/nelgiriyewithana/emotions

[25] N. A. Alam, "Phishing Email Dataset," Kaggle. Accessed: Nov. 13, 2025. [Online]. Available: https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset