Integrasi Fitur TF-IDF dan Fitur Struktural URL dengan Pendekatan Late Fusion untuk Deteksi Spam Berbahasa Indonesia

Woro Isti Rahayu, Mifthahul Fauziah*, Rina Helmina, Fatia Amalia Maresti
Program Studi Sains Data, Universitas Logistik dan Bisnis Internasional (ULBI)

Received: 25 August 2026 Accepted: 22 September 2026 Published: 1 December 2026

*mifthahulfauziah@gmail.com

Kata Kunci: : Spam, TF-IDF, Support Vector Machine, Late Fusion, Fitur Struktural URL.

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Penyebaran spam berbahasa Indonesia yang mengandung indikasi penipuan dan tautan phishing menjadi permasalahan dalam keamanan komunikasi digital. Pesan tersebut memiliki karakteristik linguistik dan dapat mengandung tautan dengan struktur mencurigakan. Penelitian ini bertujuan mengevaluasi integrasi informasi tekstual dan karakteristik URL untuk deteksi spam berbahasa Indonesia menggunakan pendekatan Late Fusion. Fitur tekstual direpresentasikan menggunakan Term Frequency–Inverse Document Frequency (TF-IDF) dan diklasifikasikan menggunakan Support Vector Machine (SVM), sedangkan informasi URL direpresentasikan melalui fitur struktural dan diklasifikasikan menggunakan LinearSVC. Keluaran probabilitas kedua model diintegrasikan menggunakan *meta-classifier* Logistic Regression. Evaluasi dilakukan melalui tiga skenario, yaitu model berbasis teks, URL, dan Late Fusion menggunakan *accuracy, precision, recall, F1-score,* dan *AUC*. Model berbasis teks memperoleh *accuracy* 97,05% dan *AUC* 98,77%, sedangkan model berbasis URL memperoleh *accuracy* 93,54% dan *AUC* 98,30%. Pada data validasi, Late Fusion menghasilkan *F1-score* 0,9788, sedangkan pada data uji 0,9663, lebih rendah dibandingkan model berbasis teks yang mencapai 0,9705. Pengujian pada data lintas platform menghasilkan *accuracy* dan *F1-score* sebesar 100%, sedangkan pada data SMS eksternal diperoleh *accuracy* 68,75% dan *F1-score* 68,63% (masing-masing n=16); nilai sempurna pada data lintas platform masih bersifat indikatif mengingat ukuran sampel yang terbatas. Hasil menunjukkan bahwa Late Fusion mampu mengintegrasikan informasi teks dan URL dalam satu mekanisme klasifikasi, tetapi penambahan fitur URL belum meningkatkan performa secara konsisten dibandingkan fitur tekstual secara mandiri.

PENDAHULUAN

Penetrasi internet di Indonesia mencapai 80,66% pada 2025[1]. Seiring manfaatnya, media digital juga disalahgunakan untuk menyebarkan spam yang tidak terbatas pada promosi, tetapi juga penipuan dan aktivitas berbahaya lainnya[2]. Data APJII mencatat penipuan daring sebagai kasus keamanan digital yang paling banyak dilaporkan (22,12% pada 2025)[1] ; survei cfds UGM menunjukkan 98,3% responden pernah menerima pesan penipuan digital, dengan SMS/telepon sebagai salah satu media utama [3]; dan OJK bersama IASC mencatat kerugian lebih dari Rp2,6 triliun hingga Mei 2025[4]. Kondisi ini menuntut mekanisme otomatis untuk mengidentifikasi pesan yang berindikasi penipuan.

Pesan penipuan umumnya memanfaatkan rekayasa sosial melalui bahasa persuasif dan dorongan untuk membuka tautan[5]. URL phishing sendiri dapat memiliki ciri struktural tertentu, seperti panjang URL, penggunaan subdomain, karakter khusus, dan pola alamat tidak lazim [6]. Dengan demikian, teks pesan

dan struktur URL berpotensi menjadi sumber informasi yang saling melengkapi.

Penelitian mengevaluasi integrasi fitur teks dan fitur struktural URL dengan Late Fusion untuk deteksi spam berbahasa Indonesia melalui tiga skenario: (1) model teks (TF-IDF + SVM), (2) model URL (fitur struktural+linearsvc), dan (3) Late Fusion dengan Logistic Regression sebagai *meta-classifier*. Selain membandingkan performa, penelitian menganalisis kapan integrasi memberi nilai tambah dan kapan tidak, serta mengimplementasikan model dalam prototipe aplikasi web.

TINJAUAN LITERATUR

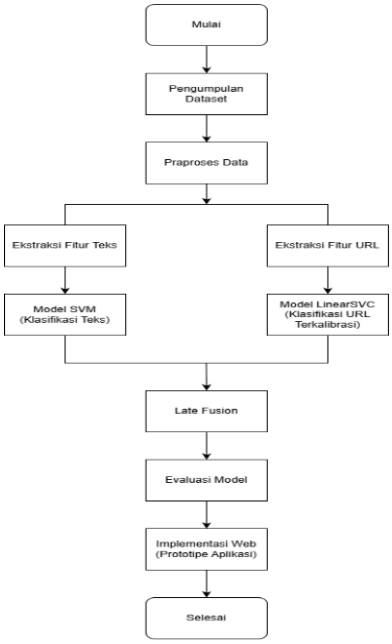
Pratama dan Ardianto mengklasifikasikan SMS berbahasa Indonesia ke dalam kategori normal, penipuan, dan promo menggunakan TF-IDF dan SVM linear, dengan *accuracy* 92%[7]. Pada sisi URL, Gurning dkk. mengklasifikasikan URL menjadi *normal, suspicious*, dan *malicious* berdasarkan panjang URL, subdomain, HTTPS, karakter khusus, dan kata kunci mencurigakan dengan pendekatan *rule-based* dan API URLHaus[8]. Silcilia dkk. membandingkan Random Forest, Decision Tree, dan Naive Bayes pada fitur URL; Random Forest terbaik dengan *accuracy* 80,75% [9]. Kedua studi menunjukkan fitur struktural URL informatif untuk mengenali tautan berbahaya.

Penelitian *fusion* menunjukkan keluaran beberapa model dapat digabungkan setelah tiap sumber diproses independen[10][11]. Pereira dkk. menemukan Late Fusion lebih menguntungkan daripada Early Fusion saat data latih terbatas[10], sedangkan Alahmadi dkk. melaporkan peningkatan performa dari penggabungan dua model independen[11].

Berdasarkan penelitian yang ditinjau, integrasi isi pesan dan karakteristik URL dengan Late Fusion untuk deteksi spam berbahasa Indonesia belum banyak dikaji. Penelitian ini mengisi celah tersebut dan mengkarakterisasi kondisi ketika integrasi memberi nilai tambah atau belum mengungguli modalitas yang lebih kuat.

METODE PENELITIAN

Penelitian menggunakan pendekatan kuantitatif eksperimental untuk mengevaluasi integrasi fitur tekstual berbasis TF-IDF dan fitur struktural URL dalam mendeteksi SMS spam berbahasa Indonesia menggunakan pendekatan Late Fusion. Metodologi penelitian dilaksanakan melalui serangkaian tahapan sistematis meliputi pengumpulan data, integrasi dan pelabelan, praproses, ekstraksi fitur, pemodelan per cabang, penggabungan model melalui late fusion, evaluasi, dan implementasi prototipe aplikasi web, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

1. Dataset dan praproses
   Dua dataset diproses terpisah: pesan SMS berbahasa Indonesia (Spam/Non-Spam) dari dua sumber, dan dataset URL untuk melatih model karakteristik struktural tautan. Dataset pesan digabung, dideduplikasi, dan pesan kurang dari tiga kata dihapus; label dibinarisasi (1 = Spam, 0 = Non-Spam). Praproses cabang teks meliputi normalisasi spasi, *case folding*, penggantian URL dan nomor telepon dengan token khusus, penghapusan tanda baca, angka, dan *stopword*. Pada cabang URL dilakukan normalisasi format, penghapusan duplikat, dan *stratified sampling* untuk menyeimbangkan kelas Phishing dan Non-Phishing.
2. Ekstraksi fitur
   Fitur diekstraksi terpisah di tiap cabang agar independen sebelum fusion. Bobot TF-IDF dihitung dengan [12]:
   $$\mathrm{TFIDF(t,d)} = \mathrm{TF(t,d)} \times \log\left(\frac{\mathrm{N}}{\mathrm{df(t)}}\right) \quad \text{.....(1)}$$
   dengan TF(t,d) frekuensi kata t pada dokumen d, N jumlah dokumen, dan df(t) jumlah dokumen yang memuat t. Fitur teks berupa TF-IDF unigram dan bigram ditambah tujuh fitur numerik (panjang pesan, jumlah kata, jumlah karakter, rata-rata panjang kata, jumlah kata kunci spam, keberadaan URL, keberadaan nomor telepon). Fitur URL berupa TF-IDF *character n-gram* (3–5 karakter) ditambah dua belas fitur struktural: panjang URL, jumlah titik, jumlah karakter khusus, HTTPS, alamat IP, kedalaman URL, simbol "@", jumlah subdomain, panjang domain, indikator URL legitimate, layanan penyingkat URL, dan rasio digit.
   Karena kedua dataset berasal dari sumber berbeda, pasangan data untuk fusion dibentuk saat *runtime*: URL dalam pesan dideteksi dengan ekspresi reguler, diekstraksi, dinormalisasi, lalu diproses model URL (yang dilatih independen pada dataset URL publik) untuk menghasilkan probabilitas phishing. Model URL berfungsi sebagai *feature extractor* probabilistik, bukan dilatih ulang pada URL dari dataset pesan.
3. Pemodelan per cabang
   Cabang teks memakai SVM dengan *Grid Search* (kernel linear dan RBF), dengan fungsi keputusan:
   $$\mathrm{f(x)} = \mathrm{sign}(\mathrm{w^T x + b}) \quad \text{.....(2)}$$
   Cabang URL memakai LinearSVC agar efisien pada fitur berdimensi tinggi dan *sparse*. Karena tidak menghasilkan probabilitas bawaan, dilakukan kalibrasi *Platt Scaling* [13] :
   $$P(y=1|x) = \frac{1}{1+\exp(A \cdot f(x)+B)} \quad \text{.....(3)}$$
   dengan A dan B dipelajari melalui *cross-validation*.
4. Late Fusion berbasis stacking
   Label Spam (model teks) dan Phishing (model URL) merepresentasikan konsep berbeda, tetapi keduanya menilai dua karakteristik pada pesan yang sama; pesan penipuan sering memakai tautan untuk mengarahkan korban ke serangan phishing [14]. Karena kategori Spam bersifat heterogen (promosi, penipuan, dan lainnya), probabilitas phishing hanya dipakai sebagai sinyal tambahan, bukan bukti langsung bahwa pesan adalah spam. Probabilitas kedua cabang menjadi *meta-features* bagi Logistic Regression, sehingga fungsi penggabungan dipelajari dari data, bukan ditetapkan manual [10], [11] :
   $$X_{meta} = \left[p_{teks}, p_{url}, 1_{has_{url}}\right] \quad \text{.....(4)}$$
   $$P(y=1|X_{meta}) = \frac{1}{1+\exp\left(-(w^T X_{meta}+b)\right)} \quad \text{.....(5)}$$
   Mekanisme *gating* diterapkan: jika pesan tidak memuat URL, *meta-classifier* dilewati dan probabilitas akhir langsung memakai p_teks dengan *threshold* 0,50 (konsisten dengan Skenario 1). Jika URL ada, *meta-classifier* aktif dengan *threshold* τ yang dioptimasi pada *validation set* berdasarkan *F1-score* maksimum. Dari 37 pesan ber-URL pada data validasi diperoleh τ = 0,2103; jumlah sampel yang kecil ini menjadi sumber ketidakpastian interpretasi. Keputusan akhir: $\hat{y}$ = 1 jika P_final ≥ τ, selain itu $\hat{y}$ = 0.
   Untuk mencegah *data leakage*, probabilitas teks pada data latih dibangun dengan *5-fold cross-validation (out-of-fold)*; *meta-classifier* dilatih pada subset data latih ber-URL dengan P(teks) *out-*

of-fold dan P(URL) dari model URL independen, sesuai prinsip *stacked generalization* [15], [16]. Threshold dioptimasi pada *validation set* yang tidak dipakai melatih *meta-classifier*, sehingga data uji tetap independen.

5. Evaluasi Model
   Evaluasi dilakukan pada tiga skenario, yaitu model berbasis teks saja (Skenario 1), model berbasis URL saja (Skenario 2), dan model integrasi Late Fusion (Skenario 3), menggunakan metrik *Accuracy, Precision, Recall, F1-score,* dan *AUC*:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (9)$$

6. Implementasi Prototipe Aplikasi Web
   Model Late Fusion terbaik hasil evaluasi selanjutnya diimplementasikan ke dalam prototipe aplikasi web sebagai bentuk pengujian penerapan model secara praktis. Prototipe dibangun menggunakan framework Flask sebagai antarmuka bagi pengguna untuk menguji deteksi pesan secara langsung berdasarkan isi teks dan URL yang disertakan dalam pesan.

HASIL PENELITIAN

Penelitian menghasilkan tiga model klasifikasi yang dibangun dan dievaluasi secara bertahap, yaitu model berbasis teks (Skenario 1), model berbasis URL (Skenario 2), dan model integrasi Late Fusion (Skenario 3). Ketiga skenario tersebut digunakan untuk mengevaluasi performa masing-masing sumber informasi serta mengetahui pengaruh integrasi fitur teks dan URL terhadap kinerja deteksi spam. Setelah proses integrasi dan pelabelan, dilakukan analisis distribusi dataset pesan untuk mengetahui komposisi data yang digunakan pada cabang teks. Distribusi dataset tersebut ditampilkan pada Gambar 2.

```
Data valid untuk training: 1576
  Spam: 668 | Non-Spam: 908
Train: 1103 | Validation: 236 | Test: 237

Dimensi fitur S1: 2007 (rasio fitur:sampel = 1.82)
```

Gambar 2. Distribusi dataset pesan setelah integrasi dan pelabelan

Berdasarkan Gambar 2, dataset pesan yang digunakan berjumlah 1.576 data, yang terdiri atas 1.103 data latih, 236 data validasi, dan 237 data uji. Pembagian data dilakukan menggunakan stratified sampling untuk mempertahankan proporsi kelas pada masing-masing subset. Setelah proses ekstraksi fitur, diperoleh representasi fitur akhir sebanyak 2.007 dimensi.

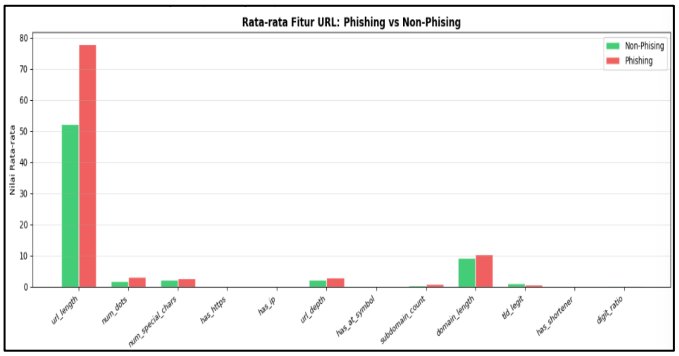
Setelah proses praproses dan sampling dilakukan, dataset URL selanjutnya digunakan untuk membangun model pada cabang URL. Tahap ini bertujuan memastikan bahwa data yang digunakan memiliki jumlah sampel yang memadai pada masing-masing tahap pengujian serta representasi fitur yang sesuai dengan kebutuhan klasifikasi. Distribusi dataset URL yang digunakan pada proses pelatihan dan pengujian ditampilkan pada Gambar 3.

```
Training URL: 20,000 | Testing URL: 5,000
Jumlah fitur S2: 5012
```

Gambar 3. Distribusi dataset URL setelah praproses dan sampling

Berdasarkan Gambar 3, dataset URL yang digunakan terdiri atas 20.000 data untuk pelatihan dan 5.000 data untuk pengujian. Setelah proses ekstraksi dan penggabungan fitur TF-IDF berbasis karakter dengan fitur struktural URL, diperoleh representasi fitur akhir sebanyak 5.012 dimensi. Pembagian tersebut digunakan sebagai dasar dalam melatih dan mengevaluasi model pada cabang URL.

Untuk melihat karakteristik data yang digunakan pada cabang URL, dilakukan analisis deskriptif terhadap fitur-fitur struktural yang diekstraksi dari URL. Analisis ini bertujuan mengidentifikasi perbedaan karakteristik antara URL pada kelas phishing dan non-phishing sebelum digunakan dalam proses klasifikasi. Perbandingan nilai rata-rata fitur struktural URL pada kedua kelas ditampilkan pada Gambar 4.



Gambar 4. Perbandingan rata-rata fitur struktural URL antara kelas phishing dan non-phishing

Berdasarkan Gambar 4, terdapat perbedaan nilai rata-rata pada beberapa fitur struktural antara URL phishing dan non-phishing. Fitur panjang URL (*url length*) menunjukkan perbedaan yang paling menonjol, dengan rata-rata URL phishing mendekati 80 karakter, sedangkan URL non-phishing memiliki rata-rata yang lebih rendah. Perbedaan juga terlihat pada beberapa karakteristik lainnya, seperti jumlah titik, karakter khusus, dan karakteristik domain. Perbedaan nilai rata-rata tersebut menunjukkan bahwa fitur struktural URL memiliki pola yang dapat dimanfaatkan oleh model untuk membedakan URL phishing dan non-phishing.

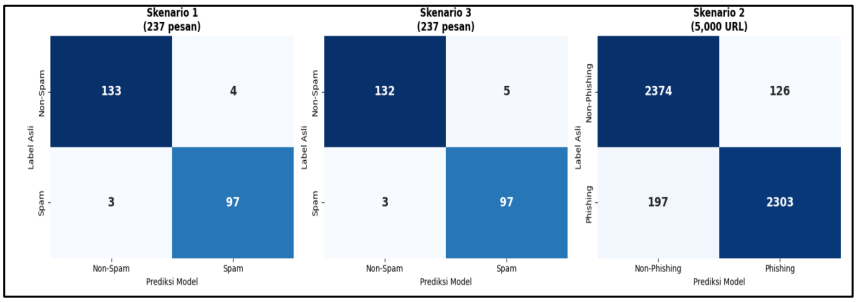
Setelah masing-masing cabang model dibangun secara independen, tahap selanjutnya dilakukan evaluasi untuk membandingkan performa model berbasis teks, model berbasis URL, dan model integrasi Late Fusion. Perbandingan ini diperlukan untuk mengetahui performa masing-masing sumber informasi sekaligus mengevaluasi apakah penggabungan kedua sumber informasi dapat memberikan peningkatan dibandingkan model tunggal. Evaluasi dilakukan menggunakan metrik *F1-score, accuracy, precision*, *recall,* dan *AUC*. Hasil perbandingan ketiga skenario disajikan pada Tabel 1.

Tabel 1. Perbandingan Metrik Evaluasi Antar Skenario

Skenario	Train F1	Val F1	Test F1	Gap	Accuracy	Precision	Recall	AUC
S1 : TF-IDF Teks + SVM	0,9946	0,9788	0,9705	0,0241	0,9705	0,9705	0,9705	0,9877
S2 : TF-IDF URL + LinearSVC	0,9644	—	0,9354	0,0291	0,9354	0,9358	0,9354	0,9830
S3 : Late Fusion (Stacking)	0,9946	0,9788	0,9663	0,0283	0,9662	0,9665	0,9662	0,9841

Berdasarkan Tabel 1, Skenario 1 yang menggunakan fitur TF-IDF teks dan SVM menghasilkan performa terbaik dengan *F1-score* sebesar 0,9705 dan AUC sebesar 0,9877. Skenario 2 yang menggunakan fitur TF-IDF URL dan LinearSVC menghasilkan *F1-score* sebesar 0,9354, sedangkan Skenario 3 menggunakan Late Fusion menghasilkan *F1-score* sebesar 0,9663. Dengan demikian, pada data pengujian, model berbasis teks masih menunjukkan performa yang lebih tinggi dibandingkan model berbasis URL maupun model Late Fusion. Hasil tersebut menunjukkan bahwa penambahan informasi URL melalui Late Fusion belum memberikan peningkatan performa dibandingkan penggunaan fitur tekstual secara mandiri. Kondisi ini dapat berkaitan dengan perbedaan kekuatan sinyal prediktif antara kedua modalitas, di mana model teks memiliki performa yang lebih tinggi daripada model URL. Pada data validasi, Skenario 3 menghasilkan *F1-score* sebesar 0,9788, identik dengan Skenario 1, karena pada threshold 0,50 kedua skenario menghasilkan keputusan klasifikasi yang sama pada mayoritas data validasi yang didominasi pesan tanpa URL. Dengan demikian, nilai F1 validasi tersebut mencerminkan kesetaraan keputusan pada tahap ini dan bukan merupakan indikasi keunggulan khusus dari mekanisme Late Fusion dibandingkan model teks tunggal.

Untuk melihat lebih rinci jenis kesalahan klasifikasi yang dihasilkan oleh masing-masing skenario, dilakukan analisis menggunakan confusion matrix pada data uji. Confusion matrix digunakan untuk menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar maupun yang mengalami kesalahan klasifikasi pada setiap kelas. Hasil confusion matrix untuk Skenario 1, Skenario 3, dan Skenario 2 ditunjukkan pada Gambar 5.



Gambar 5. Confusion matrix pada data uji

Gambar 5 menunjukkan confusion matrix pada data uji untuk Skenario 1, Skenario 3, dan Skenario 2. Pada Skenario 1, model teks mengklasifikasikan 133 pesan Non-Spam dan 97 pesan Spam dengan benar, sementara terdapat 4 pesan Non-Spam yang diprediksi sebagai Spam dan 3 pesan Spam yang diprediksi sebagai Non-Spam. Skenario 3 menghasilkan pola yang relatif serupa, dengan 132 pesan Non-Spam dan 97 pesan Spam diklasifikasikan dengan benar, serta 5 pesan Non-Spam dan 3 pesan Spam yang salah diklasifikasikan. Hal ini menunjukkan bahwa perubahan keputusan antara model teks dan Late Fusion pada data uji relatif kecil. Sementara itu, pada Skenario 2, sebanyak 2.374 URL Non-Phishing dan 2.303 URL Phishing diklasifikasikan dengan benar, sedangkan 126 URL Non-Phishing diprediksi sebagai Phishing dan 197 URL Phishing diprediksi sebagai Non-Phishing. Temuan tersebut konsisten dengan metrik evaluasi pada Tabel 1, dengan Skenario 3 menghasilkan F1-score 0,9663 dan Skenario 1 sebesar 0,9705, sedangkan Skenario 2 menghasilkan F1-score 0,9354.

Untuk mengetahui bagaimana meta-classifier mengombinasikan keluaran dari kedua cabang model, selanjutnya dianalisis koefisien yang dihasilkan oleh Logistic Regression pada tahap Late Fusion. Analisis koefisien digunakan untuk melihat pengaruh relatif dari probabilitas keluaran model teks, probabilitas keluaran model URL, serta indikator keberadaan URL terhadap keputusan akhir *meta-classifier*. Nilai koefisien yang diperoleh ditampilkan pada Gambar 6.

```
Meta-feature out-of-fold dibangun untuk 1103 pesan train.
Baris ber-URL pada train: 180 (Spam=177, Non-Spam=3)
  Koef P(teks) : 5.2343
  Koef P(url)  : 3.8047
  Intercept    : -6.6840
```

Gambar 6. Koefisien *meta-classifier* Late Fusion Berbasis Stacking

Berdasarkan Gambar 6, *meta-classifier* menghasilkan koefisien sebesar 5,2343 untuk P(teks) dan 3,8047 untuk P(URL), dengan intercept sebesar −6,6840. Kedua koefisien bernilai positif, menunjukkan bahwa peningkatan probabilitas spam dari cabang teks maupun probabilitas phishing dari cabang URL berkaitan dengan peningkatan probabilitas keluaran *meta-classifier*. Kedua sinyal tersebut digunakan dalam pembentukan keputusan meta-classifier, dengan koefisien numerik P(teks) yang lebih besar dibandingkan P(URL).

Untuk mengetahui lebih lanjut pengaruh mekanisme Late Fusion berdasarkan ketersediaan informasi URL, data uji selanjutnya dianalisis dengan membagi pesan menjadi dua subset, yaitu pesan yang tidak mengandung URL dan pesan yang mengandung URL. Analisis ini dilakukan karena mekanisme *gating* pada sistem menentukan jalur klasifikasi berdasarkan keberadaan URL. Pada pesan tanpa URL, prediksi menggunakan keluaran model teks secara langsung, sedangkan pada pesan ber-URL, keluaran model teks dan model URL digunakan dalam proses Late Fusion. Perbandingan *F1-score* kedua skenario pada masing-masing subset ditampilkan pada Tabel 2.

Tabel 2. Perbandingan F1-score skenario 1 dan skenario 3 berdasarkan keberadaan URL

Subset	F1 Skenario 1	F1 Skenario 3	Selisih
Pesan Tanpa URL (203 Pesan)	0,9656	0,9607	-0,0049
Pesan Ber-URL (34 Pesan)	1,0000	1,0000	0,0000

Berdasarkan Tabel 2, pada subset pesan tanpa URL, Skenario 1 menghasilkan *F1-score* sebesar 0,9656, sedangkan Skenario 3 menghasilkan *F1-score* sebesar 0,9607. Selisih yang relatif kecil ini muncul karena label akhir Skenario 1 menggunakan fungsi keputusan asli SVM (*predict*), sedangkan pada subset tanpa URL, Skenario 3 menurunkan keputusan dari probabilitas hasil kalibrasi Platt Scaling (predict_proba

≥ 0,50), kedua aturan keputusan ini tidak selalu identik pada kasus yang berada dekat batas keputusan. Perbedaan ini bersifat implementasional dan tidak memengaruhi perbandingan performa utama antar-skenario, karena setiap skenario tetap dievaluasi menggunakan aturan keputusannya sendiri secara konsisten. Mekanisme gating tetap memastikan proses Late Fusion (*meta-classifier*) tidak diterapkan pada subset tersebut.

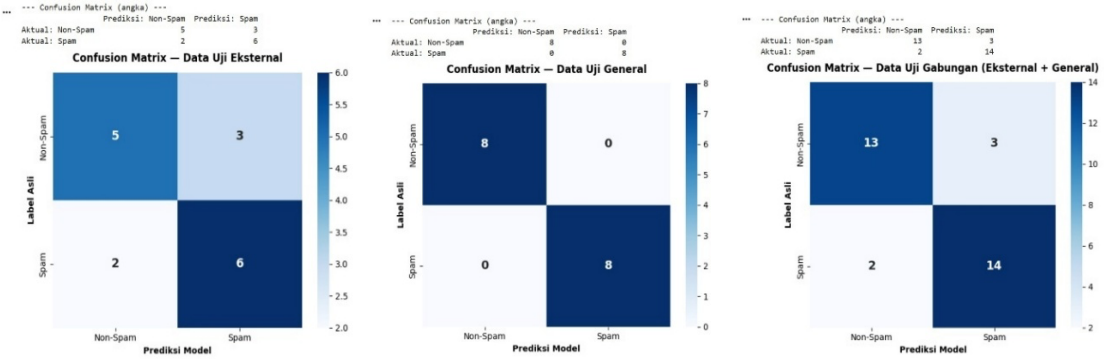
Pengujian tambahan dilakukan menggunakan data uji manual yang berada di luar dataset pelatihan untuk memperoleh gambaran awal mengenai kemampuan generalisasi model terhadap data yang tidak digunakan selama proses pelatihan. Data tersebut terdiri atas pesan SMS sebagai data eksternal dan pesan lintas platform yang berasal dari beberapa media komunikasi, seperti WhatsApp, Telegram, dan email sebagai data general. Hasil pengujian pada kedua kelompok data tersebut disajikan pada Tabel 3.

Tabel 3. Hasil pengujian pada data uji manual

Sumber Data Uji	N	Accuracy	Precision	Recall	F1
Eksternal (Pesan SMS)	16	0,6875	0,690	0,688	0,6863
General (Pesan Lintas Platform)	16	1,0000	1,0000	1,0000	1,0000
Gabungan	32	0,8438	0,8451	0,8438	0,8436

Berdasarkan Tabel 3, terlihat adanya perbedaan performa yang cukup besar antara kedua kelompok data uji manual. Model mencapai *accuracy* sebesar 100% pada data lintas platform, sedangkan pada data SMS eksternal *accuracy* yang diperoleh hanya sebesar 68,75%. Perbedaan tersebut menunjukkan bahwa performa model dapat berubah ketika diterapkan pada data yang memiliki karakteristik berbeda dari data pelatihan. Namun, hasil tersebut masih memiliki keterbatasan dalam menggambarkan kemampuan generalisasi model karena masing-masing kelompok hanya terdiri atas 16 pesan. Dengan ukuran sampel n=16, satu kesalahan klasifikasi dapat mengubah *accuracy* sebesar 6,25 poin persentase. Oleh karena itu, hasil pengujian manual belum dapat menggambarkan performa generalisasi model secara stabil dan lebih sesuai digunakan sebagai indikasi awal.

Untuk membantu memahami distribusi kesalahan klasifikasi pada masing-masing kelompok data uji manual, terutama karena nilai *accuracy, precision, recall*, dan *F1-score* pada beberapa kelompok bernilai hampir sama, confusion matrix ketiga kelompok data uji tersebut ditampilkan pada gambar 7.



Gambar 7. Confusion Matrix Pada Data Uji Manual

Berdasarkan gambar 7, pada data uji General seluruh 16 pesan diklasifikasikan dengan benar tanpa kesalahan (FP = 0, FN = 0), yang menjelaskan *accuracy* dan *F1-score* sebesar 100% pada kelompok tersebut. Pada data uji Eksternal, model menghasilkan 3 kesalahan *false positive* dan 2 kesalahan *false negative* dari 16 pesan, yang menunjukkan bahwa kesalahan klasifikasi terjadi pada kedua arah, yaitu pesan Non-Spam yang salah diprediksi sebagai Spam maupun sebaliknya. Pada data gabungan, total kesalahan sebanyak 5 dari 32 pesan (3 FP dan 2 FN) konsisten dengan *accuracy* gabungan sebesar 84,38%. Distribusi tersebut menunjukkan bahwa kesalahan klasifikasi pada data uji Eksternal tidak terkonsentrasi pada satu kelas tertentu, melainkan tersebar pada kedua kelas.

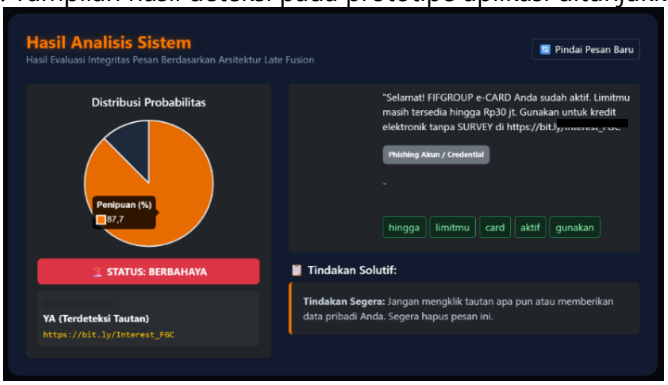
Untuk mengilustrasikan penerapan mekanisme Late Fusion yang dijelaskan pada Persamaan (4) dan (5) pada data riil, Tabel 4 menyajikan contoh alur deteksi pesan beserta probabilitas keluaran masing-masing cabang model.

Tabel 4. Contoh Alur Keputusan Late Fusion

No.	Pesan	P(teks)	P(url)	P(final)	τ	Keputusan
1	Selamat! FIFGROUP e-CARD Anda sudah aktif. Limitmu masih tersedia hingga Rp30 jt. Gunakan untuk kredit elektronik tanpa SURVEY di https://bit.ly/lxxxx	0,9795	0,6573	0,7198	0,2103	Spam
2	Halo semuanya 😊 Mohon bantuannya untuk like dan membagikan postingan kelompok kami sebagai bagian dari penilaian AAS. Setiap like yang diberikan sangat berarti. Terima kasih banyak! 🙏 https://www.instagram.com/p/xxxxxxxxxx/	0,3357	0,4069	0,0330	0,2103	Non-Spam
3	Halo, jadi kita ketemuan jam 7 malam ya di cafe biasa	0,0083	-	0,0083	0,5	Non-Spam

Berdasarkan Tabel 4, pada Contoh 1, pesan yang menyertakan tautan pemendek URL (bit.ly) menghasilkan probabilitas spam dari model teks yang tinggi (0,9795) serta probabilitas phishing dari model URL sebesar 0,6573, sehingga P(final) yang dihasilkan oleh *meta-classifier* sebesar 0,7198, melampaui threshold τ = 0,2103, dan pesan diklasifikasikan sebagai Spam. Pada Contoh 2, pesan yang menyertakan tautan Instagram menghasilkan probabilitas spam dari model teks sebesar 0,3357 dan probabilitas phishing dari model URL sebesar 0,4069; kombinasi keduanya menghasilkan P(final) sebesar 0,0330, yang masih berada di bawah threshold τ = 0,2103, sehingga pesan diklasifikasikan sebagai Non-Spam. Pada Contoh 3, pesan tidak mengandung URL sehingga mekanisme *gating* pada Persamaan (5) melewati tahap *meta-classifier* dan probabilitas akhir ditentukan langsung dari keluaran model teks (0,0083) menggunakan threshold default 0,50, sehingga pesan diklasifikasikan sebagai Non-Spam. Ketiga contoh tersebut menunjukkan bahwa keputusan akhir Late Fusion bergantung pada keberadaan URL pada pesan, apabila URL ditemukan, kedua probabilitas dikombinasikan melalui *meta-classifier* dan dievaluasi terhadap threshold τ = 0,2103 hasil optimasi. Sebaliknya, apabila URL tidak ditemukan, keputusan sepenuhnya mengikuti model teks dengan threshold default 0,50, konsisten dengan mekanisme yang dijelaskan pada Persamaan (5).

Sebagai bentuk implementasi dari model yang telah dikembangkan, penelitian ini juga menghasilkan prototipe aplikasi web yang bernama SATRIA (Sistem Analisis Teks & Radar Identifikasi Ancaman), prototipe aplikasi ini digunakan untuk menerapkan proses deteksi pesan secara langsung. Prototipe ini dirancang untuk menampilkan hasil klasifikasi beserta informasi pendukung yang dapat membantu pengguna memahami hasil deteksi. Tampilan hasil deteksi pada prototipe aplikasi ditunjukkan pada Gambar 8.



Gambar 8. Tampilan hasil deteksi pada prototipe aplikasi

Gambar 8 menunjukkan tampilan hasil deteksi pada prototipe aplikasi SATRIA. Sistem menampilkan distribusi probabilitas hasil klasifikasi dalam bentuk diagram lingkaran serta status keamanan pesan yang terdeteksi. Selain hasil klasifikasi, sistem juga menampilkan kategori penjelasan berdasarkan kata kunci yang teridentifikasi, seperti "Phishing Akun/Credential", serta informasi mengenai URL yang terdeteksi pada pesan. Informasi tersebut dilengkapi dengan analisis singkat dan rekomendasi tindakan solutif bagi pengguna. Tampilan tersebut menunjukkan penerapan hasil klasifikasi ke dalam bentuk yang lebih mudah digunakan, sehingga keluaran model tidak hanya berupa label klasifikasi, tetapi juga disertai informasi pendukung yang dapat membantu pengguna memahami hasil deteksi.

PEMBAHASAN

Hasil penelitian menunjukkan bahwa integrasi fitur teks dan URL melalui pendekatan Late Fusion tidak selalu menghasilkan peningkatan performa dibandingkan model berbasis teks tunggal. Skenario 3 memperoleh *F1-score* sebesar 0,9663, sedikit lebih rendah dibandingkan Skenario 1 yang mencapai 0,9705, penurunan tersebut sepenuhnya berasal dari subset pesan tanpa URL (Tabel 2), sedangkan pada subset pesan ber-URL kedua skenario justru menghasilkan *F1-score* yang identik (1,0000). Temuan ini sejalan dengan analisis teoretis Pereira dkk. [10], yang menunjukkan bahwa manfaat Late Fusion bergantung pada keseimbangan kekuatan informasi antar-modalitas. Ketika salah satu modalitas memiliki separabilitas yang jauh lebih rendah dibandingkan modalitas lainnya, penggabungan skor dapat menghasilkan performa yang lebih rendah dibandingkan penggunaan modalitas yang lebih kuat secara mandiri. Temuan serupa juga dilaporkan pada domain klasifikasi multimodal secara umum, di mana jaringan unimodal terbaik dapat mengungguli jaringan multimodal ketika salah satu modalitas memiliki sinyal yang jauh lebih lemah [17] . Pada penelitian ini, performa cabang URL (*F1-score* = 0,9354) berada di bawah performa cabang teks (*F1-score* = 0,9705), sehingga ketidakseimbangan kekuatan sinyal tersebut dapat menjadi salah satu faktor yang menjelaskan hasil Late Fusion pada Tabel 1 dan Tabel 2.

Perbedaan performa Skenario 3 juga perlu dilihat dari proses penentuan threshold yang dioptimasi menggunakan 37 pesan ber-URL pada data validasi. Ukuran sampel tersebut relatif terbatas sehingga threshold yang diperoleh berpotensi kurang stabil dan dapat memengaruhi keputusan pada data uji. Kondisi ini sejalan dengan temuan pada arsitektur late fusion lain seperti RG-Hybrid, di mana kontribusi cabang tambahan terhadap peningkatan performa cenderung kecil ketika model utama sudah memiliki performa yang tinggi sehingga ruang peningkatan yang tersedia menjadi terbatas [11]. Dengan demikian, performa model integrasi tidak hanya dipengaruhi oleh kualitas kedua modalitas, tetapi juga oleh kestabilan proses penentuan threshold. Pada subset pesan tanpa URL, Skenario 1 memperoleh *F1-score* sebesar 0,9656, sedangkan Skenario 3 memperoleh *F1-score* sebesar 0,9607 (Tabel 2), sehingga penerapan mekanisme gating menghasilkan penurunan F1-score sebesar 0,0049 dibandingkan model teks tunggal. Meskipun penurunannya relatif kecil, hasil ini menunjukkan bahwa integrasi melalui Late Fusion tidak selalu setara dengan model teks tunggal pada pesan tanpa URL. Pada subset pesan yang mengandung URL, kedua skenario memperoleh F1-score yang sama, yaitu 1,0000, sehingga integrasi sinyal tekstual dan URL tidak menghasilkan perbedaan kinerja pada subset tersebut. Dengan demikian, manfaat fusion dalam penelitian ini lebih tepat dipahami sebagai kontribusi yang bersifat kondisional dan bergantung pada ketersediaan URL, bukan sebagai peningkatan atau pemeliharaan performa yang bersifat absolut.

Penelitian memiliki mekanisme kategorisasi berbasis kata kunci pada prototipe SATRIA yang digunakan untuk memberikan informasi penjelas, seperti penipuan berhadiah, phishing akun, dan promosi/iklan, dan lain-lain. Kategorisasi tersebut bersifat heuristik dan belum didasarkan pada anotasi *ground-truth* yang tervalidasi, sehingga tidak digunakan sebagai label dalam pelatihan maupun evaluasi model. Oleh karena itu, penelitian ini belum dapat mengevaluasi kontribusi sinyal URL secara kuantitatif pada masing-masing motif spam. Pengembangan anotasi motif spam yang tervalidasi dan analisis berdasarkan subkategori tersebut menjadi salah satu arah penelitian selanjutnya.

Pengujian pada data uji manual menunjukkan kesenjangan performa antara data lintas platform dengan *accuracy* 100% dan data pesan SMS dengan *accuracy* 68,75%. Kesenjangan tersebut mengindikasikan adanya kemungkinan bahwa karakteristik linguistik pada data pelatihan belum sepenuhnya mewakili variasi gaya bahasa SMS di luar distribusi pelatihan. Performa sempurna pada data lintas platform juga perlu dilihat dalam konteks ukuran sampel yang terbatas, yaitu n=16. Dengan ukuran tersebut, satu kesalahan klasifikasi akan mengubah *accuracy* sebesar 6,25 poin persentase. Oleh karena itu, hasil pengujian manual ini belum dapat digunakan untuk
menyimpulkan kemampuan generalisasi model secara menyeluruh dan lebih tepat diposisikan sebagai pengujian awal di luar distribusi data pelatihan.

Dari sisi implementasi, prototipe aplikasi SATRIA menghadirkan mekanisme penjelasan (*explainability*) berupa visualisasi probabilitas dan rekomendasi tindakan solutif. Pendekatan ini relevan dengan penelitian sebelumnya yang menunjukkan bahwa transparansi pada sistem keamanan berbasis kecerdasan buatan dapat meningkatkan kepercayaan pengguna terhadap keputusan otomatis [18]. Penggunaan peringatan

yang disertai penjelasan alasan dan rekomendasi tindakan juga sejalan dengan temuan penelitian sebelumnya mengenai efektivitas penjelasan dalam membantu pengguna memahami dan menghindari ancaman [19]. Kategori eksplanasi berbasis kata kunci yang ditampilkan turut membantu memberikan konteks terhadap jenis ancaman yang terdeteksi, sejalan dengan klasifikasi jenis penipuan konsumen ke dalam kategori penipuan berbasis peluang (*opportunity-based*) dan penipuan phishing [20].

Secara keseluruhan, temuan menunjukkan bahwa integrasi fitur teks dan URL melalui pendekatan Late Fusion belum secara konsisten mengungguli model berbasis teks tunggal pada penelitian ini, terutama ketika terdapat kesenjangan kekuatan sinyal antara kedua modalitas. Koefisien *meta-classifier* menunjukkan bahwa sinyal tekstual lebih dominan daripada sinyal URL, sementara koefisien URL yang tetap positif menunjukkan bahwa informasi URL masih dipertimbangkan sebagai sinyal pelengkap. Namun, kontribusi tersebut belum cukup kuat untuk meningkatkan *F1-score* di atas model teks pada data uji. Di sisi lain, mekanisme gating memastikan bahwa penurunan performa pada pesan tanpa URL tetap kecil (selisih *F1-score* 0,0049), jauh lebih kecil dibandingkan potensi degradasi apabila Late Fusion diterapkan secara penuh tanpa mekanisme kondisional tersebut. Dengan demikian, kontribusi utama penelitian ini bukan pada peningkatan performa absolut melalui Late Fusion, melainkan pada karakterisasi empiris mengenai kondisi ketika integrasi dua modalitas memberikan nilai tambah atau belum memberikan peningkatan pada deteksi spam berbahasa Indonesia. Temuan ini memberikan dasar untuk penelitian selanjutnya melalui penambahan volume data ber-URL, evaluasi sensitivitas threshold, serta pengujian pada data lintas platform yang lebih besar dan beragam.

KESIMPULAN

Penelitian berhasil membangun dan mengevaluasi tiga skenario klasifikasi spam berbahasa Indonesia, yaitu model berbasis teks, model berbasis URL, dan model integrasi Late Fusion berbasis *meta-classifier* stacking. Hasil pengujian menunjukkan bahwa model berbasis teks memberikan performa terbaik pada data uji dengan *F1-score* sebesar 0,9705, sedangkan model Late Fusion menghasilkan *F1-score* sebesar 0,9663. Analisis koefisien *meta-classifier* menunjukkan bahwa kedua sinyal memberikan kontribusi positif terhadap keputusan klasifikasi, dengan koefisien P(teks) sebesar 5,2343 dan P(URL) sebesar 3,8047. Koefisien P(teks) yang lebih besar menunjukkan bahwa sinyal tekstual memiliki kontribusi yang lebih kuat dalam *meta-classifier* dibandingkan sinyal URL pada konfigurasi penelitian ini, sejalan dengan performa Skenario 1 yang sedikit lebih tinggi dibandingkan Skenario 3 pada pengujian keseluruhan (0,9705 vs 0,9663).

Analisis berdasarkan keberadaan URL menunjukkan bahwa mekanisme *gating* menjaga performa Skenario 3 tetap dekat dengan Skenario 1 pada pesan tanpa URL, meskipun terdapat selisih *F1-score* sebesar 0,0049. Namun, pada pesan yang mengandung URL, Late Fusion menghasilkan *F1-score* sebesar 1,0000, setara dengan model teks yang juga mencapai 1,0000. Dengan demikian, integrasi fitur URL melalui Late Fusion belum menunjukkan peningkatan performa yang konsisten dibandingkan penggunaan fitur tekstual secara mandiri. Pengujian manual pada data di luar dataset pelatihan menunjukkan perbedaan performa antara data lintas platform dengan *accuracy* 100% dan SMS eksternal dengan *accuracy* 68,75%. Namun, hasil tersebut masih bersifat indikatif karena masing-masing kelompok hanya terdiri atas 16 pesan.

Secara keseluruhan, kontribusi utama penelitian ini terletak pada karakterisasi empiris penggunaan Late Fusion untuk menggabungkan sinyal teks dan URL dalam deteksi spam berbahasa Indonesia. Pada penelitian ini, ketika sinyal tekstual lebih kuat daripada sinyal URL, integrasi kedua modalitas belum menghasilkan peningkatan performa. Di sisi lain, mekanisme *gating* membatasi besarnya penurunan performa pada pesan tanpa URL, meskipun penurunan kecil (selisih *F1-score* 0,0049) tetap teramati dibandingkan model teks tunggal. Temuan ini memberikan dasar bagi penelitian selanjutnya untuk menggunakan data ber-URL yang lebih besar, threshold yang lebih stabil, serta data uji lintas platform yang lebih beragam.

REFERENCES

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, "Survei Penetrasi Internet dan Perilaku Penggunaan Internet 2025." Diakses: 10 Juli 2026. [Daring]. Tersedia pada: https://survei.apjii.or.id/

[2] O. Abayomi-Alli, S. Misra, dan A. Abayomi-Alli, "A deep learning method for automatic <scp>SMS</scp> spam classification: Performance of learning algorithms on indigenous dataset," *Concurr. Comput.*, vol. 34, no. 17, Agu 2022, doi: 10.1002/cpe.6989.

[3] N. Kurnia *dkk.*, *Penipuan Digital di Indonesia: Modus, Medium, dan Rekomendasi*. Sleman, Indonesia: Program Studi Magister Ilmu Komunikasi, Fakultas Ilmu Sosial dan Ilmu Politik, Universitas Gadjah Mada, 2022. Diakses: 10 Juli 2026. [Daring]. Tersedia pada: https://pr2media.or.id/wp-content/uploads/2022/10/PDF-Penipuan-Digital-di-Indonesia-Modus-Medium-dan-Rekomendasi_.pdf?

[4] E. Wahyuni, "OJK terima 128 ribu aduan scamming, total kerugian Rp2,6 T." Diakses: 10 Juli 2026. [Daring]. Tersedia pada: https://wantimpres.go.id/id/newsflows/ojk-terima-128-ribu-aduan-scamming-total-kerugian-rp26-t/

[5] E. Haryadi, D. Wijayanti, dan E. Chandra, "Pelatihan pencegahan SMS phishing (smishing) pada warga Perumahan Papan Mas Tambun," *Prawara Jurnal Abdimas*, vol. 5, no. 1, hlm. 1–6, Feb 2026, Diakses: 11 Juli 2026. [Daring]. Tersedia pada: https://www.jurnalilmiah.id/index.php/abdimas/article/view/258/186

[6] F. Alwan, R. A. Pasha, E. P. Kaspandiri, dan M. D. A. Fariz, "Analisis serangan phishing dan upaya pencegahannya pada sistem informasi," *JIKUM (Jurnal Ilmu Komputer)*, vol. 2, no. 2, Nov 2026, Diakses: 11 Juli 2026. [Daring]. Tersedia pada: https://journal.dcircle.org/index.php/jikum/article/download/177/130

[7] M. N. Pratama dan D. P. Ardianto, "Klasifikasi pesan SMS menggunakan metode TF-IDF dan support vector machine," 2026. Diakses: 11 Juli 2026. [Daring]. Tersedia pada: https://proceeding.unpkediri.ac.id/index.php/stains/article/view/9094/5952

[8] Y. S. M. G. Yeremia, Dedy Kiswanto, Leni Karmila Daulay, dan Aurela Khoiri Nasution, "SISTEM KLASIFIKASI URL UNTUK DETEKSI SITUS BERBAHAYA BERBASIS ANALISIS FITUR JARINGAN," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, no. 1, Jan 2026, doi: 10.23960/jitet.v14i1.8186.

[9] Silcilia, N. P. Salsabila, dan T. Apriani, "Analisis Performa Random Forest, Decision Tree, dan Naive Bayes untuk Deteksi Link Phishing Berbasis Fitur URL," *ROUTERS: Jurnal Sistem dan Teknologi Informasi*, vol. 4, no. 1, hlm. 13–20, Feb 2026, doi: 10.25181/rt.v5i1.4330.

[10] L. M. Pereira, A. Salazar, dan L. Vergara, "A Comparative Analysis of Early and Late Fusion for the Multimodal Two-Class Problem," *IEEE Access*, vol. 11, hlm. 84283–84300, 2023, doi: 10.1109/ACCESS.2023.3296098.

[11] K. Alahmadi, S. Alharbi, dan X. Wang, "RG-Hybrid: pure late fusion of RoBERTa and a graph transformer for robust, interpretable sentiment classification," *Int. J. Data Sci. Anal.*, vol. 22, no. 1, hlm. 137, Des 2026, doi: 10.1007/s41060-026-01112-x.

[12] D. Septiani dan I. Isabela, "Analisis term frequency inverse document frequency (TF-IDF) dalam temu kembali informasi pada dokumen teks," *Sintesia: Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 1, no. 2, 2023, Diakses: 12 Juli 2026. [Daring]. Tersedia pada: https://journal.unj.ac.id/unj/index.php/SINTESIA/article/view/39364

[13] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," dalam *Advances in Large Margin Classifiers*, A. Smola, P. J. Bartlett, B. Schölkopf, dan D. Schuurmans, Ed., Cambridge, MA: MIT Press, 1999.

[14] I. Tomičić, "Social Engineering Aspects of Email Phishing: an Overview and Taxonomy," dalam *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, IEEE, Mei 2023, hlm. 1201–1207. doi: 10.23919/MIPRO57284.2023.10159691.

[15] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, hlm. 241–259, Jan 1992, doi: 10.1016/S0893-6080(05)80023-1.

[16] K. M. Ting dan I. H. Witten, "Issues in Stacked Generalization," *Journal of Artificial Intelligence Research*, vol. 10, hlm. 271–289, Mei 1999, doi: 10.1613/jair.594.

[17] W. Wang, D. Tran, dan M. Feiszli, "What Makes Training Multi-Modal Classification Networks Hard?,"

2020, hlm. 12695–12705.
[18] N. Capuano, G. Fenza, V. Loia, dan C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, hlm. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.
[19] G. Desolda, J. Aneke, C. Ardito, R. Lanzilotti, dan M. F. Costabile, "Explanations in warning dialogs to help users defend against phishing attacks," *Int. J. Hum. Comput. Stud.*, vol. 176, hlm. 103056, Agu 2023, doi: 10.1016/j.ijhcs.2023.103056.
[20] M. DeLiema, Y. Li, dan G. Mottola, "Correlates of responding to and becoming victimized by fraud: Examining risk factors by scam type," *Int. J. Consum. Stud.*, vol. 47, no. 3, hlm. 1042–1059, Mei 2023, doi: 10.1111/ijcs.12886.