e- ISSN: 2828–299X

Sistem Deteksi Spam Email Menggunakan Machine Learning

Aldo Aditya Saputra[1], Rahmat Nanda Eka Saputra [2]
[1,2]Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri
Email: *[1]aldoadityasaputra70@gmail.com, [2]rnandajr758@gmail.com

Abstrak – *Perkembangan teknologi informasi mendorong peningkatan penggunaan email sebagai media komunikasi digital, namun di sisi lain juga meningkatkan jumlah spam email yang berpotensi mengganggu kenyamanan dan keamanan pengguna. Oleh karena itu, diperlukan sistem deteksi spam email yang mampu bekerja secara otomatis dan akurat. Penelitian ini bertujuan untuk mengembangkan sistem deteksi spam email menggunakan pendekatan machine learning dengan algoritma Regresi Logistik. Metode penelitian meliputi pengumpulan dataset email berbahasa Indonesia, preprocessing teks, ekstraksi fitur menggunakan Term Frequency–Inverse Document Frequency (TF-IDF), pelatihan dan pengujian model, serta evaluasi kinerja sistem. Pengujian dilakukan menggunakan pembagian data 80% untuk data latih dan 20% untuk data uji. Hasil evaluasi menunjukkan bahwa model yang dikembangkan mampu mencapai akurasi sebesar 97,38%, dengan nilai precision, recall, dan F1-score yang seimbang pada kedua kelas spam dan non-spam. Selain itu, model diimplementasikan dalam bentuk aplikasi berbasis Streamlit sehingga dapat digunakan secara interaktif. Hasil penelitian ini menunjukkan bahwa kombinasi TF-IDF dan Regresi Logistik efektif untuk mendeteksi spam email berbahasa Indonesia dan berpotensi diterapkan sebagai sistem penyaringan email otomatis.*

Kata Kunci *— deteksi spam, email, machine learning, regresi logistik, TF-IDF*

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat telah mendorong peningkatan penggunaan email sebagai media komunikasi digital[1]. Namun, seiring dengan tingginya intensitas penggunaan email, berbagai permasalahan mulai muncul, salah satunya adalah spam email. Spam email merupakan pesan yang dikirim secara massal tanpa persetujuan penerima dan sering kali berisi konten tidak relevan, promosi agresif, hingga ancaman seperti *phishing*. Keberadaan spam tidak hanya mengganggu kenyamanan pengguna, tetapi juga dapat menimbulkan potensi risiko terkait keamanan data dan informasi pribadi[2].

Kebutuhan akan sistem penyaringan email yang lebih cerdas semakin mendesak karena metode manual tidak lagi efisien dalam menangani volume pesan yang besar. Pendekatan berbasis machine learning menjadi solusi yang banyak digunakan karena mampu menganalisis pola teks secara otomatis dan mendeteksi indikasi spam dengan tingkat akurasi yang lebih tinggi. Dalam pemrosesan teks, teknik *TF-IDF (Term Frequency–Inverse Document Frequency)* sering dimanfaatkan untuk mengekstraksi fitur dari dokumen karena kemampuannya memberikan bobot signifikan pada kata-kata yang dianggap penting[3].

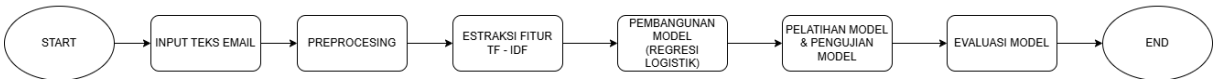
Metode *Regresi Logistik*, sebagai salah satu algoritma klasifikasi biner, juga banyak digunakan dalam permasalahan seperti pemisahan email spam dan non-spam. Algoritma ini bekerja dengan menghitung probabilitas suatu data termasuk dalam kategori tertentu berdasarkan fitur yang telah diekstraksi. Kombinasi *TF-IDF* dan *Regresi Logistik* dikenal efektif, sederhana, dan mampu memberikan performa yang baik dalam klasifikasi berbasis teks[4].

Berdasarkan kondisi tersebut, penelitian ini bertujuan mengembangkan sistem deteksi spam email yang memanfaatkan *TF-IDF* sebagai metode ekstraksi fitur dan *Regresi Logistik* sebagai algoritma klasifikasi. Sistem ini diharapkan dapat meningkatkan efisiensi proses penyaringan email serta memberikan akurasi yang optimal dalam membedakan pesan spam dan pesan valid. Selain itu, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan teknologi keamanan email yang lebih andal dan mudah diimplementasikan[5].

2. METODE PENELITIAN

Pada bagian ini dijelaskan tahapan penelitian yang mencakup proses pengumpulan data, preprocessing teks, ekstraksi fitur, pembangunan model, pelatihan model, serta evaluasi model. Alur umum proses penelitian ditampilkan pada Gambar 2.1, yang menggambarkan rangkaian tahapan dimulai dari input data hingga proses evaluasi. Penelitian ini diawali dengan memuat dataset email yang telah diberi label sebagai spam dan non-spam (ham). Dataset tersebut kemudian diproses melalui tahapan pembersihan teks untuk memastikan bahwa seluruh data berada dalam format yang sesuai sebelum digunakan dalam proses pembelajaran mesin. Tahapan ini

dilakukan secara sistematis agar menghasilkan model klasifikasi spam yang akurat dan stabil. Seluruh alur tersebut ditampilkan secara runtut pada Gambar 2.1.



Gambar 1 Alur Proses Penelitian

2.1 Input Data/Teks Email

Tahap pertama adalah memuat dataset pesan berbahasa Indonesia yang sudah diberi label spam dan ham. Dataset diambil dari website yaitu Kaggle. Dataset yang digunakan pada penelitian ini memiliki dua kolom utama, yaitu Kategori (spam/ham) dan Pesan (teks pesan). Total data yang digunakan sebanyak 1.143 data, dengan distribusi 574 spam dan 569 ham. Data disusun agar siap diproses pada tahapan berikutnya.

2.2 Preprocesing

Tahap kedua adalah *Preprocessing Data*, yang dilakukan untuk menyiapkan teks sebelum diolah oleh algoritma machine learning. *Preprocessing* merupakan tahap penting karena data mentah biasanya mengandung banyak noise yang dapat mengurangi akurasi model. Proses *preprocessing* yang dilakukan dalam penelitian ini meliputi:

1) *Case Folding,* yaitu mengubah seluruh karakter menjadi huruf kecil agar konsisten;
2) *Tokenization,* yaitu memecah teks menjadi unit kata;
3) *Stopword Removal,* yaitu menghapus kata-kata umum seperti “yang”, “dan”, “atau” yang tidak memberikan kontribusi pada klasifikasi;
4) *Stemming,* yaitu mengubah kata ke bentuk dasarnya menggunakan stemmer Bahasa Indonesia; serta
5) *Cleaning,* yaitu menghapus angka, tanda baca, URL, dan karakter yang tidak relevan. Hasil dari tahap ini adalah teks yang telah bersih dan siap diolah lebih lanjut.

2.3 Ekstraksi Fitur menggunakan *TF-IDF*

Tahap ketiga adalah Ekstraksi Fitur menggunakan *TF-IDF (Term Frequency–Inverse Document Frequency)*. *TF-IDF* digunakan untuk mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma klasifikasi[6]. Metode ini memberikan bobot pada setiap kata berdasarkan frekuensinya dalam sebuah dokumen serta tingkat kekhasannya dalam keseluruhan dataset. Kata-kata yang sering muncul dalam email tertentu tetapi jarang muncul di email lain akan diberikan bobot lebih tinggi. Proses ekstraksi fitur ini menghasilkan matriks *TF-IDF* yang digunakan sebagai masukan pada tahap pelatihan model.

2.4 Pembangunan Model

Tahap selanjutnya adalah Perancangan Model Klasifikasi, yaitu menggunakan algoritma *Regresi Logistik*. *Regresi Logistik* merupakan algoritma klasifikasi biner yang memodelkan probabilitas sebuah data termasuk kelas spam atau ham berdasarkan fitur yang telah diekstraksi[7]. Pemilihan algoritma ini dilatarbelakangi oleh keunggulannya yang sederhana, efisien, dan memiliki performa yang sangat baik untuk data berbasis teks. Arsitektur model terdiri dari fungsi aktivasi sigmoid sebagai pengubah keluaran menjadi nilai probabilitas[8]. Berikut adalah fungsi sigmoid :

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \quad \text{..... 1)}$$

2.5 Pelatihan dan Pengujian Model

Tahap berikutnya adalah Pelatihan Model (Training), di mana dataset dibagi menjadi dua bagian utama yaitu 80% untuk pelatihan dan 20% untuk pengujian. Pada tahap ini, data pelatihan digunakan untuk menghasilkan parameter terbaik pada algoritma *Regresi Logistik* melalui proses *gradient descent*. Selama proses pelatihan, model mempelajari pola-pola karakteristik pada email spam, termasuk penggunaan kata-kata tertentu, struktur kalimat, serta frasa yang sering muncul. Pelatihan berlangsung hingga model mencapai tingkat konvergensi atau kondisi terbaik berdasarkan nilai loss. Setelah proses pelatihan selesai, dilakukan Pengujian Model *(Testing)* menggunakan data uji yang belum pernah dilihat oleh model sebelumnya untuk mengukur kemampuan generalisasinya dalam mengklasifikasikan email nyata. Hasil pengujian kemudian dianalisis untuk menilai apakah model mampu bekerja secara efektif dalam kondisi sebenarnya.

2.6 Evaluasi Model

Tahap terakhir adalah Evaluasi Model, yang dilakukan menggunakan beberapa metrik performa, yaitu akurasi, *precision*, *recall*, dan *F1-score*. Akurasi digunakan untuk mengetahui persentase prediksi yang benar, precision mengukur ketepatan model dalam mendeteksi spam, recall mengukur kemampuan model menemukan seluruh email spam, dan *F1-score* digunakan untuk mendapatkan keseimbangan antara *precision* dan *recall*[9]. Selain itu, penelitian ini juga menggunakan teknik *cross-validation* untuk memastikan bahwa hasil evaluasi stabil dan tidak *overfitting*. Seluruh hasil evaluasi ini kemudian dianalisis pada bagian hasil dan pembahasan.

1) Confusion Matrix
   Confusion matrix untuk klasifikasi biner terdiri dari empat komponen:
   a) *TP (True Positive):* pesan spam yang diprediksi spam
   b) *TN (True Negative):* pesan ham yang diprediksi ham
   c) *FP (False Positive):* pesan ham yang diprediksi spam
   d) *FN (False Negative):* pesan spam yang diprediksi ham
2) Akurasi *(Accuracy)*
   Akurasi menunjukkan proporsi prediksi yang benar terhadap seluruh data uji.
   Berikut adalah rumus akurasi:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{..... 2)}$$

3) Presisi
   Presisi mengukur ketepatan model saat memprediksi kelas spam. Semakin tinggi precision, semakin kecil kemungkinan pesan ham dianggap spam.
   Berikut adalah rumus presisi:

$$Precision = \frac{TP}{TP + FP} \qquad \text{..... 3)}$$

4) Recall (*Sensitivity*)
   Recall mengukur kemampuan model menemukan seluruh spam yang ada. Semakin tinggi recall, semakin sedikit spam yang lolos sebagai ham.
   Berikut adalah rumus recall:

$$Recall = \frac{TP}{TP + FN} \qquad \text{..... 4)}$$

5) F1-Score
   F1-score adalah rata-rata harmonik dari precision dan recall, digunakan untuk menilai performa yang seimbang[10].
   Berikut adalah rumus F1-Score:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \text{..... 5)}$$

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data dan Skenario Pengujian

Dataset yang digunakan berjumlah 1.143 pesan yang terdiri dari 574 spam dan 569 ham. Pengujian dilakukan dengan pembagian data 80% data latih dan 20% data uji sesuai kode pelatihan model.

Model dibangun menggunakan pipeline *TF-IDF* dan *Regresi Logistik*. *TF-IDF* menggunakan konfigurasi ngram_range (1,2), max_df 0.95, dan min_df 2. Sementara Regresi Logistik menggunakan max_iter 1000 dan solver liblinear.

Tabel 1 Distribusi Dataset

Kelas	Jumlah
Spam	574
Ham	569
Total	**1143**

3.2 Hasil Evaluasi Kinerja Model

Evaluasi kinerja model deteksi spam email dilakukan menggunakan data uji sebanyak 229 data, yang terdiri dari 111 data non-spam (kelas 0) dan 118 data spam (kelas 1). Model yang digunakan adalah *Regresi Logistik* dengan fitur *TF-IDF*, dan evaluasi dilakukan menggunakan metrik *akurasi, precision, recall,* dan *F1-score*.

Hasil pengujian menunjukkan bahwa model memperoleh nilai akurasi sebesar 97,38%. Nilai ini menunjukkan bahwa sebagian besar pesan email pada data uji berhasil diklasifikasikan dengan benar ke dalam kategori spam maupun non-spam.

Tabel 2 Hasil Evaluasi Kinerja Model Deteksi Spam Email

Kelas	Precision	Recall	F1-Score	Support
Non-Spam (0)	0,96	0,98	0,97	111
Spam (1)	0,98	0,97	0,97	118
Akurasi			0,97	229
Macro Average	0,97	0,97	0,97	229
Weighted Average	0,97	0,97	0,97	229

Berdasarkan Tabel 2, dapat dijelaskan sebagai berikut:

a) *Precision* kelas spam sebesar 0,98 menunjukkan bahwa 98% pesan yang diprediksi sebagai spam benar-benar merupakan spam. Hal ini menandakan bahwa model sangat baik dalam menghindari kesalahan klasifikasi pesan penting sebagai spam *(false positive)*.
b) *Recall* kelas spam sebesar 0,97 menunjukkan bahwa 97% dari seluruh pesan spam berhasil terdeteksi oleh sistem. Dengan nilai *recall* yang tinggi, risiko spam lolos ke *inbox* pengguna dapat diminimalkan.
c) *F1-score* sebesar 0,97 pada kedua kelas menunjukkan keseimbangan yang baik antara *precision* dan *recall*, sehingga model tidak bias terhadap salah satu kelas.

3.3 Hasil Evaluasi Menggunakan *Confusion Matrix*

Hasil prediksi pada data uji dirangkum menggunakan *confusion matrix*. Pada penelitian ini, kelas spam dianggap sebagai kelas positif.

Tabel 3 Confusion Matrix (Data Uji)

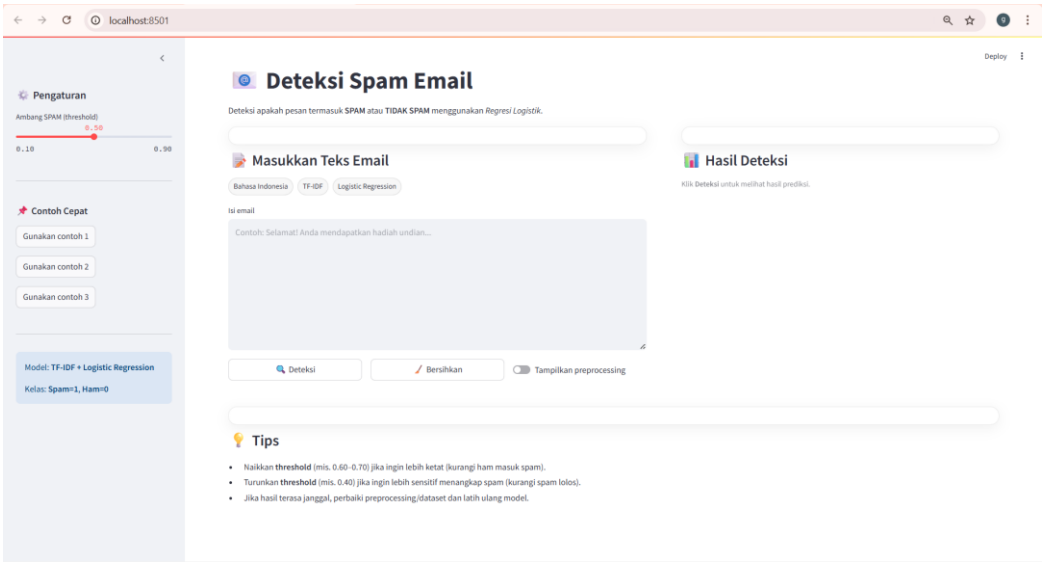
Aktual \ Prediksi	Ham	Spam
Ham	110	4
Spam	4	111

Interpretasi hasil:

a) *True Negative (TN)* = 110: ham terklasifikasi benar sebagai ham.
b) *False Positive (FP)* = 4: ham salah terklasifikasi sebagai spam.
c) *False Negative (FN)* = 4: spam salah terklasifikasi sebagai ham.
d) *True Positive (TP)* = 111: spam terklasifikasi benar sebagai spam.

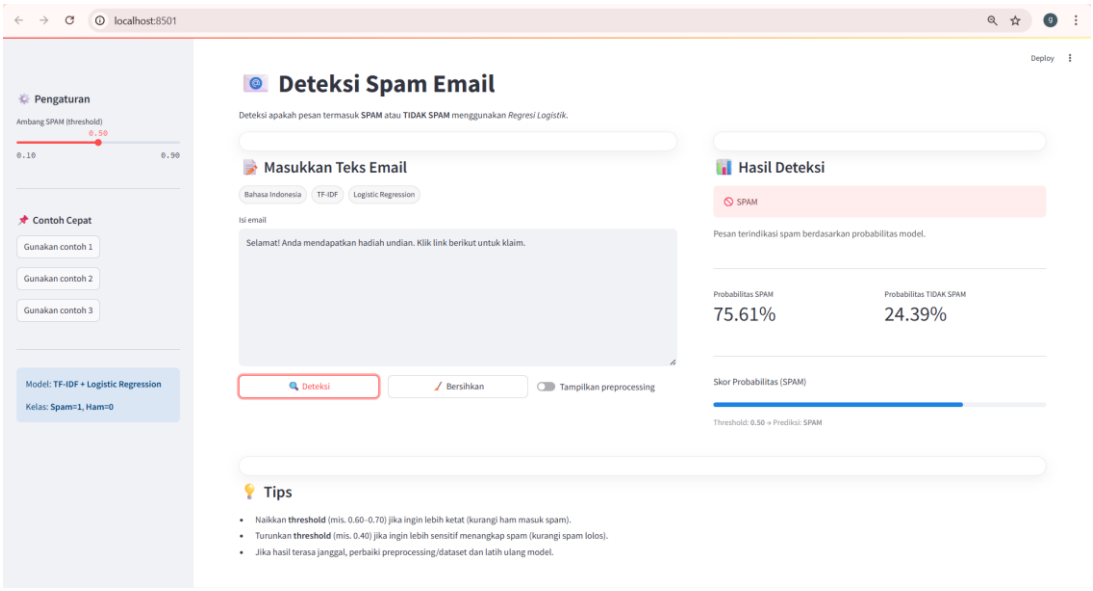
Secara umum, kesalahan model relatif kecil. Kesalahan FP biasanya muncul ketika pesan ham mengandung kata/frasa yang mirip promosi, sedangkan kesalahan FN terjadi ketika pesan spam dibuat menyerupai pesan normal sehingga pola katanya tidak cukup "kuat" terdeteksi sebagai spam.

3.4 Hasil Implementasi Aplikasi Streamlit



Gambar 2 Tampilan UI implementasi Aplikasi Deteksi Spam Email

Gambar 2 merupakan tampilan aplikasi Deteksi Spam Email pada gambar menunjukkan antarmuka sistem klasifikasi spam yang dirancang sederhana, informatif, dan mudah digunakan. Pada bagian kiri terdapat sidebar pengaturan yang memungkinkan pengguna menyesuaikan ambang batas *(threshold)* spam, sehingga tingkat sensitivitas deteksi dapat diatur sesuai kebutuhan. Sidebar ini juga menyediakan contoh cepat untuk membantu pengguna menguji sistem tanpa harus mengetik teks secara manual, serta informasi singkat mengenai model yang digunakan, yaitu *TF-IDF* dan *Regresi Logistik* dengan dua kelas klasifikasi (spam dan non-spam). Pada bagian utama, pengguna disediakan area Masukkan Teks Email untuk menuliskan isi pesan yang akan diuji, lengkap dengan penanda metode yang digunakan agar pengguna memahami proses klasifikasi di balik sistem. Tombol Deteksi digunakan untuk menjalankan prediksi, sementara tombol Bersihkan berfungsi menghapus teks input, dan opsi tampilkan *preprocessing* memungkinkan pengguna melihat hasil pembersihan teks sebelum diklasifikasikan. Di sisi kanan terdapat panel Hasil Deteksi yang menampilkan status prediksi apakah pesan termasuk spam atau tidak spam setelah proses dijalankan. Pada bagian bawah, aplikasi juga dilengkapi dengan tips penggunaan yang memberikan panduan interpretasi hasil dan pengaturan *threshold*. Secara keseluruhan, tampilan ini menunjukkan implementasi sistem deteksi spam email yang tidak hanya berfungsi secara teknis, tetapi juga memperhatikan aspek keterpahaman dan kemudahan interaksi pengguna.



Gambar 3 Tampilan Implementasi Hasil Deteksi Aplikasi Deteksi Spam Email

Tampilan aplikasi Deteksi Spam Email pada gambar 3 menunjukkan kondisi ketika sistem telah melakukan proses klasifikasi terhadap teks email yang dimasukkan oleh pengguna. Pada bagian kiri terdapat sidebar pengaturan yang memungkinkan pengguna mengatur nilai *threshold* spam untuk menentukan tingkat sensitivitas deteksi, menyediakan contoh cepat untuk pengujian, serta menampilkan informasi singkat mengenai model yang digunakan, yaitu *TF-IDF* dan *Regresi Logistik* dengan dua kelas keluaran spam dan non-spam. Pada bagian utama, pengguna dapat melihat teks email yang diuji pada area Masukkan Teks Email, kemudian setelah tombol Deteksi ditekan, hasil klasifikasi ditampilkan pada panel Hasil Deteksi di sisi kanan. Sistem menunjukkan bahwa pesan termasuk SPAM, disertai dengan informasi probabilitas spam sebesar 75,61% dan probabilitas tidak spam sebesar 24,39%, serta visualisasi berupa progress bar yang menggambarkan tingkat keyakinan model. Di bagian bawah, aplikasi juga menyediakan tips penggunaan untuk membantu pengguna memahami dan menyesuaikan pengaturan *threshold* sesuai kebutuhan. Secara keseluruhan, antarmuka ini menampilkan sistem deteksi spam yang interaktif, informatif, dan transparan, karena tidak hanya memberikan hasil klasifikasi, tetapi juga menyajikan nilai probabilitas sebagai dasar pengambilan keputusan.

Keunggulan Implementasi Aplikasi:

a) Pengguna dapat mengatur *threshold* pada sidebar untuk menyesuaikan tingkat sensitivitas spam.
b) Sistem menampilkan probabilitas spam dan tidak spam, sehingga hasil keputusan model lebih transparan.
c) Tersedia contoh cepat dan fitur tampilan *preprocessing* untuk membantu pengguna memahami proses klasifikasi.

4. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan sebagai berikut:

1. Sistem deteksi spam email berbasis *machine learning* menggunakan algoritma *Regresi Logistik* dan fitur *TF-IDF* berhasil dikembangkan dan diimplementasikan dengan baik.
2. Model yang dibangun mampu mengklasifikasikan pesan email ke dalam kategori spam dan non-spam dengan akurasi sebesar 97,38%, serta nilai *precision*, *recall*, dan *F1-score* yang seimbang pada kedua kelas.
3. Pendekatan *TF-IDF* efektif dalam merepresentasikan teks email berbahasa Indonesia, sedangkan *Regresi Logistik* mampu memanfaatkan fitur tersebut untuk melakukan klasifikasi secara akurat dan efisien.
4. Implementasi model ke dalam aplikasi berbasis Streamlit memberikan kemudahan penggunaan serta transparansi hasil prediksi melalui penyajian probabilitas.
5. Keterbatasan penelitian ini terletak pada jumlah dan variasi dataset yang masih terbatas, sehingga performa model dapat ditingkatkan dengan data yang lebih beragam pada penelitian selanjutnya.

SARAN

Berdasarkan keterbatasan penelitian yang telah dilakukan, saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Menggunakan dataset email dengan jumlah yang lebih besar dan variasi bahasa yang lebih beragam agar model memiliki kemampuan generalisasi yang lebih baik.

2. Mengembangkan sistem agar dapat terintegrasi langsung dengan layanan email secara real-time sehingga dapat digunakan dalam lingkungan nyata.

DAFTAR PUSTAKA

[1] M. Danuri, “Perkembangan Dan Transformasi Teknologi Digital,” *Jurnal Ilmiah Infokam*, vol. 15, no. 2, pp. 116–123, 2019, doi: 10.53845/infokam.v15i2.178.

[2] M. Soleh and Z. Tjenreng, “Strategi Pencegahan Kebocoran Data Pelayanan Publik Di Era Digital,” *Jurnal Kajian Pemerintah: Journal of Government, Social and Politics*, vol. 11, no. 1, pp. 1–10, 2024, doi: 10.25299/jkp.2025.vol11(1).20524.

[3] D. Septiani and I. Isabela, “Term Frequency Inverse Document Frequency (Tf-Idf) Analysis in Information Retrieval in Text Documents,” *Jurnal Sistem dan Teknologi Informasi Indonesia(SINTESIA)*, vol. 1, no. 2, pp. 81–88, 2022.

[4] A. Atikah Putri, S. Agustian, and R. Abdillah, “Penerapan Metode Logistic Regression Untuk Klasifikasi Sentimen Pada Dataset Twitter Terbatas,” *Jurnal Sistem Informasi*, vol. 7, no. 1, pp. 95–107, 2025.

[5] J. Riset, B. C. Utomo, and A. A. Rahman, "Analisis Kesadaran Keamanan Data Pribadi pada Pengguna E-Wallet Analysis of Personal Data Security Awareness of DANA E-Wallet Users," vol. 8, no. 2, pp. 155–166, 2024.

[6] A. D. Meisya Putri, N. Sulistianingsih, and R. Rismayati, "JTIM : Jurnal Teknologi Informasi dan Multimedia Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF," vol. 7, no. 4, pp. 675–688, 2025.

[7] A. Nur, R. Hasanah, R. A. Krestianti, and S. Wati, "Implementasi Algoritma Regresi Logistik untuk Binary Classification dalam Spam SMS dan WhatsApp," *Inotek*, vol. 7, pp. 80–93, 2023, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/inotex/

[8] D. H. Tanjung, "Jaringan Saraf Tiruan dengan Backpropagation untuk Memprediksi Penyakit Asma," *Creative Information Technology Journal*, vol. 2, no. 1, p. 28, 2015, doi: 10.24076/citec.2014v2i1.35.

[9] R. Fauzan, A. V. Vitianingsih, D. Cahyono, A. L. Maukar, and Y. A. B. Suprio, "Application of Classification Algorithms in Machine Learning for Phishing Detection," *Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 2, pp. 531–540, 2025.

[10] T. Tukino, "Penerapan Algoritma Convolutional Neural Network Untuk Klasifikasi Sentimen Pada Layanan e-Commerce," *Jurnal Desain Dan Analisis Teknologi*, vol. 4, no. 1, pp. 44–53, 2025, doi: 10.58520/jddat.v4i1.72.