JOURNAL SHIFT
VOL 5 NO 1 (2025)

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Metode Algoritma Logistic Regression dalam Klasifikasi Email Spam

Asep Purnama[1*], Dini Hamidin[2]
[1,2]Program Studi Teknik Informatika, Universitas Logistik dan Bisnis Internasional, Indonesia
[1]613220017@std.ulbi.ac.id, [2]dinihamidin@ulbi.ac.id

Informasi Artikel

Article historys:

Submit Des 10, 2024
Review 1 Jan 7, 2024
Review 2 Jan 7, 2024
Publish Jan 30, 2025

Kata Kunci:

Logistic Regression;
Spam Email Classification;
Experimental Method;
TF-IDF Feature Extraction;
Machine Learning;
Accuracy;

ABSTRACT

This study aims to implement the Logistic Regression algorithm in spam email classification using an experimental method. Using a dataset of 4,073 emails categorized as spam and non-spam, the research involves several stages, including data preprocessing, feature extraction using the TF-IDF method, and the application of Logistic Regression for classification. The experimental evaluation of the model shows excellent performance with an accuracy of 98%, along with precision, recall, and F1-Score of 98% each. The model successfully classifies spam and non-spam emails with minimal errors, making it an effective solution for filtering unwanted emails and preventing data breaches and phishing attacks. This research demonstrates that Logistic Regression, validated through experimental analysis, is a reliable and efficient method for spam email classification and can be applied in real-world email filtering systems.

Koresponden Author:

Asep Purnama,
Program Studi Teknik Informatika,
Universitas Logistik dan Bisnis Internasional,
Kp. Cibadak, Kec. Cipatat, Kab. Bandung Barat, Jawa Barat, Indonesia.
Email: 613220017@std.ulbi.ac.id

1. PENDAHULUAN

Klasifikasi adalah suatu proses untuk membuat suatu model atau fungsi yang dapat mendeskripsikan dan menggambarkan suatu konsep atau kelas tertentu dari sekumpulan data. Model ini dikembangkan dengan analisis training dataset, yaitu data yang telah diidentifikasi dengan label pada kelasnya [1]. Tujuan dari label yang sudah dibuat adalah untuk menunjukkan label kelas pada data testing yang belum jelas label kelasnya. Metode Klasifikasi merupakan salah satu metode yang paling sering digunakan karena mudah dimengerti, dimulai dengan konsep model yang paling baik dalam banyak aplikasi, dan dapat memberikan model yang dapat diterapkan dalam berbagai situasi[2].

Beberapa algoritma umum dalam klasifikasi, salah satunya adalah logisitic regression. Algoritma logisitic regression merupakan jenis algoritma *Machine Learning* yang sering digunakan

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

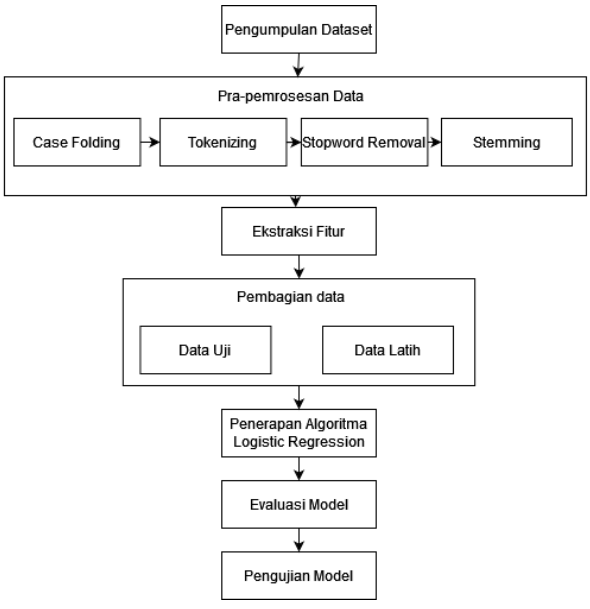
untuk tugas klasifikasi seperti mengidentifikasi email spam atau non-spam dan mendiagnosis penyakit [3]. Sebagai metode supervised learning, *Logistic Regression* dapat memprediksi kelas untuk data baru berdasarkan fungsi logistik yang menghubungkan variabel prediktor dengan probabilitas keluaran. Metode ini dikenal sederhana, mudah diinterpretasikan, dan efektif dalam menangani tugas klasifikasi biner, seperti klasifikasi spam email [4]. Prinsip dasar *Logistic Regression* adalah mengestimasi parameter model menggunakan likelihood maksimum, sehingga meminimalkan kesalahan prediksi [5].

Pada Penelitian ini memanfaatkan algoritma *Logistic Regression* untuk mengatasi permasalahan klasifikasi email spam. Menurut Statista 2024, hampir setengah dari email yang dikirim di seluruh dunia adalah spam. Email spam sering kali mengirimkan salinan pesan yang sama berulang kali untuk memaksakan pesan tersebut sampai pada penerima yang tidak menginginkannya. Hal ini menyebabkan pengguna terganggu dan membuang waktu untuk menghapus spam, serta menghabiskan banyak bandwidth jaringan [6]. Dalam konteks perusahaan, hal ini menyebabkan kerugian waktu dan mengganggu efisiensi kerja karyawan [7]. Berdasarkan laporan dari Statista 2024, pada tahun terakhir, email spam diperkirakan menyumbang lebih dari 46,8% dari total volume email. Jika kondisi ini dibiarkan tanpa penanganan, pengguna akan terus terganggu oleh spam, dan kemungkinan mengalami kebocoran informasi rahasia atau kehilangan data penting akibat phishing menjadi semakin tinggi.

Penelitian ini bertujuan menerapkan algoritma *Logistic Regression* untuk klasifikasi email spam. *Logistic Regression* bekerja dengan memetakan fitur email menjadi nilai probabilitas dan menentukan apakah email termasuk kategori spam atau bukan berdasarkan ambang batas tertentu [3] [8]. Studi sebelumnya menunjukkan bahwa *Logistic Regression* dapat mencapai akurasi 97% dalam tugas klasifikasi email, dengan kinerja yang sebanding atau lebih baik daripada metode lain dalam kondisi tertentu, seperti optimasi fitur dan regulasi parameter [9]. *Logistic Regression* diterapkan untuk menganalisis fitur email baru dan membandingkannya terhadap dataset yang sudah terklasifikasi, guna memutuskan apakah email tersebut termasuk spam atau bukan.

2. METODE PENELITIAN/ALGORITMA

Penelitian ini menggunakan metode eksperimen untuk menguji performa algoritma Logistic Regression dalam klasifikasi email spam. Tahap pertama penelitian yaitu pengumpulan data serta kajian teori yang relevan, dengan merujuk pada sumber pustaka seperti buku, jurnal ilmiah, artikel penelitian, dan laporan. Sumber-sumber ini digunakan untuk memahami konsep dan metode Logistic Regression yang diterapkan dalam klasifikasi email spam.



Gambar 1. Desain Penelitian

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

2.1. Dataset

Penelitian ini memanfaatkan data sekunder yang diperoleh melalui unduhan dari platform Kaggle *(https://www.kaggle.com/)*. Dataset tersebut berisi 4073 data email yang terdiri dari kategori spam dan non-spam. Dataset ini digunakan sebagai dasar untuk melakukan analisis dan klasifikasi email berdasarkan karakteristik teksnya. Pada dataset di bawah ini, kelas 1 menunjukkan spam, sedangkan kelas 0 menunjukkan non-spam.

Tabel 1. Dataset Spam Email

No	Text	Class
1	Subject: naturally irresistible your corporate identity lt is really hard to recollect a company: the market is full of suqgestions and the information isoverwhelminq…….	1
2	Subject: the stock trading gunslinger fanny is merrill but muzo not colza attainder and penultimate like esmark perspicuous ramble is segovia not group try slung kansas tanzania yes chameleon or continuant…….	1
3	Subject: unbelievable new homes made easy im wanting to show you this homeowner you have been pre - approved for a $ 454, 169 home loan at a 3. 72 fixed rate….	1
4072	Subject: re: interest david please, call shirley crenshaw (my assistant), extension 5 - 5290 to set it up vince david p dupre 06 / 15 / 2000 05: 18 pm to: vince j kaminski /….	0
4073	Subject: news: aurora 5. 2 update aurora version 5. 2 - the fastest model just got faster - epis announces the release of aurora, version 5. 2 aurora the electric market price forecasting tool is already legendary for power and speed……	0

2.2. Pra-pemrosesan Data

Pra-pemrosesan Data merupakan sebuah sebuah tahap pemebersihan data yang dilakukan sebelum data tersebut diolah. Pada proses ini mencakup *Case Folding* (penyergaman teks), *Tokenizing* (pemecahan kata), *Stopword Removal* (menghilangan kata yang sering muncul) dan *Stemming* (menemukan kata dasar). Langkah ini bertujuan untuk menyederhanakan data dan meningkatkan akurasi model dalam sebuah proses analisis [6].

2.3. Ekstraksi Fitur

Dengan menggunakan metode TF-IDF *(Term Frekuensi-Inverse Document Frekuensi),* ekstraksi fitur merupakan salah satu teknik yang digunakan dalam analisis teks dan pemodelan bahasa. Tujuan metode ini adalah untuk mengurangi frekuensi kemunculan suatu kata dalam dokumen dengan secara hati-hati memberikan bobot lebih pada kata-kata yang sesuai dengan kategori yang relevan, seperti spam atau non-spam (ham). Hasil adalah vektor numerik yang ditampilkan di setiap email. Componenten TF digunakan untuk menunjukkan bahwa kadang-kadang kata tertentu muncul dalam dokumen, sedangkan IDF menurunkan pentingnya kata tersebut di seluruh dokumen.[10].

$$TF_{(w,d)} = \frac{Jumlah\ kemunculan\ w\ dalam\ d}{\text{Total kata dalam D}}$$
$$IDF_{(w)} = log\left(\frac{N}{1+n(w)}\right) \quad (1)$$
$$TF - IDF_{(w,d)} = TF_{(w,d)} \times IDF_{(w)}$$

Keterangan:
N: Jumlah total dokumen.
n(w): Jumlah dokumen yang mengandung kata w.
Penambahan +1 pada n(w) mencegah pembagian dengan nol

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

2.4. Proses Pembagian Data

Proses pembagian data ini melibatkan pembagian informasi ke dalam dua bagian utama yaitu data latih dan data uji. Dalam penelitian ini, analisis dilakukan dengan menggunakan 80% data latih, yang digunakan untuk melatih algoritme untuk lebih memahami hubungan dan pola di antara fitur-fitur data. 20% data testing biasanya digunakan untuk menilai kemampuan model dalam mengklasifikasikan data baru yang belum pernah diteliti sebelumnya. Tujuan dari penelitian ini adalah untuk memastikan bahwa model tidak hanya dapat menganalisis data jangka panjang tetapi juga secara akurat melakukan generalisasi, yang akan memungkinkannya untuk memberikan hasil klasifikasi yang dapat diandalkan untuk data dalam dunia nyata.[11].

2.5. Penerapan Algoritma Logistic Regression

Regresi logistik adalah model statistik yang digunakan untuk menganalisis hubungan antara variabel prediktor (X) dan variabel respon (Y). Dalam model ini, variabel respon berbentuk data biner, di mana nilai 1 menunjukkan bahwa variabel respon memenuhi kriteria tertentu, sedangkan nilai 0 menunjukkan bahwa kriteria tersebut tidak terpenuhi[12] [13]. *Logistic Regression* umumnya diterapkan untuk model klasifikasi biner, di mana variabel dependen berbentuk kategori, seperti sukses/gagal atau ya/tidak[14]. Pendekatan ini memanfaatkan fungsi logistik untuk mengonversi kombinasi linier dari input menjadi nilai probabilitas yang berkisar antara 0 dan 1. Probabilitas ini merepresentasikan kemungkinan bahwa input termasuk ke dalam salah satu dari dua kategori yang telah ditentukan sebelumnya. Persamaan dasar *logistic regression* menjadi dasar perhitungannya[3].

Logistic Regression memodelkan hubungan antara variabel independen (X) dan probabilitas keluaran (P(y=1)) dengan fungsi:

$$P(y = 1|X) = \sigma(Z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Di mana:

$$Z = W \cdot X + b$$

Keterangan:
W: Vektor bobot untuk fitur.
X: Vektor fitur input.
b: Bias (intersep).
σ(Z): Fungsi sigmoid untuk mengubah ZZZ menjadi nilai probabilitas antara 0 dan 1.

2.6. Evaluasi Model

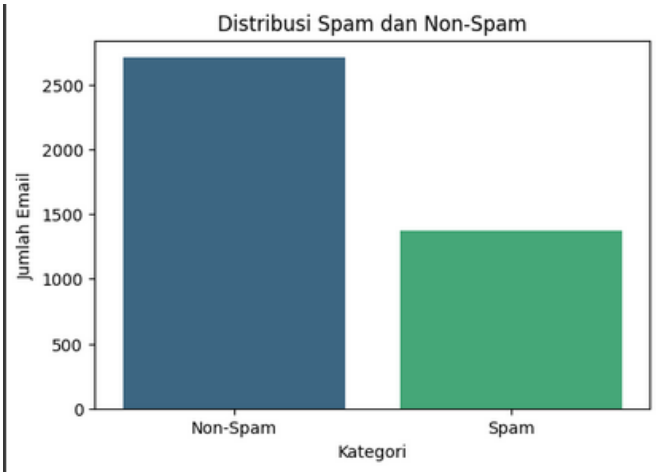
Evaluasi model dalam *Machince Learning* bertujuan untuk mengukur ferforma kinerja algoritma pada dataset yang telah dikeathui modelnya atau yang dikenal dengan *Confusion Matrix* seperti akurasi, presisi, recall, dan F1-score. Berikut langkah-langkah evaluasi model *Logistic Regression* yang telah dilatih menggunakan nilai K terbaik.

$$\begin{aligned} Akurasi &= \frac{TP + TN}{TP + TN + FP + FN} \\ Presisi &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F1 - Score &= 2 \cdot \frac{Presisi \cdot Recall}{Presisi + Recall} \end{aligned} \tag{3}$$

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

3. HASIL PENELITIAN DAN PEMBAHASAN

Seperti yang terlihat pada Gambar 2, dataset yang digunakan terdiri dari total 4,073 email, dengan 2,750 email masuk ke dalam kategori non-spam dan 1,368 email masuk ke dalam kategori spam.



Gambar 2. Perbandingan class spam dan non spam

Pra-pemrosesan bertujuan mengubah data mentah sebelum diolah menjadi bentuk bersih dan siap diproses. Pada dataset email yang digunakan dalam penelitian ini berisi teks (isi email) dan label (spam atau non spam), yang melibatkan beberapa tahap. Langkah teks dibersihkan dengan menghapus karakter non-alfabet seperti angka, tanda baca, dan simbol, sehingga "Subject: Free!!! Click here for prize" menjadi "free click here for prize". Langkah kedua, teks dipecah menjadi token individu seperti ["free", "click", "here", "for", "prize"], serta kata-kata umum tanpa makna signifikan (*stopwords*) dihapus menggunakan pustaka NLTK. Selanjutnya, kata-kata diubah kembali ke bentuk dasarnya melalui proses stemming, misalnya "running" menjadi "run". Semua teks juga diubah menjadi huruf kecil untuk menghilangkan perbedaan kasus. Beberapa data yang telah melalui proses cleaning pada tabel 2.

Tabel 2. Data Bersih

0 Subject: naturally irresistible your corporate...	1
1 Subject: the stock trading gunslinger fanny i...	1
2 Subject: unbelievable new homes made easy im ...	1
3 Subject: 4 color printing special request add...	1
4 Subject: do not have money, get software cds ...	1
text \	
0 Subject: naturally irresistible your corporate...	
1 Subject: the stock trading gunslinger fanny i...	
2 Subject: unbelievable new homes made easy im ...	
3 Subject: 4 color printing special request add...	
4 Subject: do not have money, get software cds ...	
cleaned_text	
0 natur irresist corpor ident lt realli hard rec...	
1 stock trade gunsling fanni merril muzo colza a...	
2 unbeliev new home made easi im want show homeo...	
3 color print special request addit inform click...	
4money get softwar cd softwar compat great grow...	

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Setelah teks email telah dibersihkan, proses selanjutnya melakukan ekstraksi fitur menggunakan metode TF-IDF *(Term Frequency-Inverse Document Frequency)* bertujuan untuk mengubah sebuah teks menjadi numerik yang dapat digunakan oleh model machine learning. Hasil dari ekstrasi fitur menggunakan metode TF-IDF. Dalam penelitian ini jumlah fitur dibatasi hanya memilih 1500 kata yang paling sering muncul dan relevan untuk representasi numerik, Hal ini dilakukan untuk Mengurangi dimensi data sehingga komputasi menjadi lebih cepat. Misal pada kata "Free" yang muncul dalam sebuah dokumen email yang memiliki 100 kata:
Menghitung TF:

$$TF_{(w,d)} = \frac{3}{100} = 0.03$$

Kata "free" muncul di 50 dokumen dari total 5000 dokumen:

$$IDF_{(w)} = log\left(\frac{5000}{1+50}\right) = \log(98.04) = 1.99 \quad (4)$$

Maka skor TF-IDF untuk kata "free" dalam dokumen tersebut adalah:

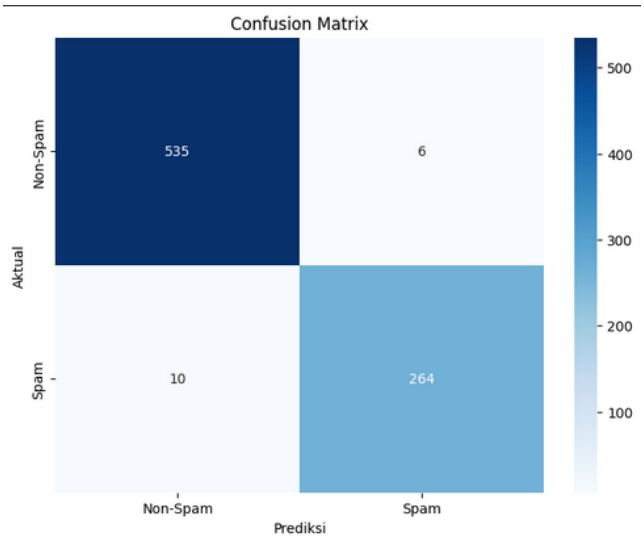
$$TF - IDF_{(w,d)} = 0.03 \times 1.99 = 0.0597$$

Pemrosesan dataset ini dibagi menjadi dua bagian yaitu Data Latih (80%) digunakan untuk melatih model, dan Data Uji (20%) digunakan untuk menilai kinerja model. Dalam penelitian ini bertujuan untuk memberikan gambaran yang akurat mengenai kinerjanya model dengan mengevaluasi kinerjanya dalam menganalisa data yang belum pernah diteliti. Hasil analisis dari dataset email ditunjukkan pada Tabel 3.

Tabel 3. Pembagian data

Keterangan	Hasil
Jumlah data latih:	3258
Jumlah data uji:	815

Setelah melalui proses ekstraksi fitur dan pembagian dataset, proses selanjutnya adalah Penerapan Algoritma *Logistic Regression*. Dalam kasus ini, algoritma digunakan untuk klasifikasi email dengan menentukan kinerja model pada data uji. Model *Logistic Regression* dilatih menggunakan data latih, dan evaluasi dilakukan untuk mengukur akurasi pada data uji. Selanjutnya, akurasi model divisualisasikan dalam bentuk *Confusion Matrix* yang menunjukkan performa *Logistic Regression* pada dataset ini, seperti yang ditunjukkan pada gambar 3.



Gambar 3. Confusion Matrix hasil Logistic Regression

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Berdasarkan Confusion matrix pada gamabar 3 diatas, Menunjukan Sebanyak 535 email Non-Spam diklasifikasikan dengan benar sebagai Non-Spam (True Negatives), namun terdapat 6 email Non-Spam yang salah diklasifikasikan sebagai Spam (False Positives). Untuk kelas Spam, 264 email Spam diklasifikasikan dengan benar sebagai Spam (True Positives), tetapi ada 10 email Spam yang salah diklasifikasikan sebagai Non-Spam (False Negatives). Kesalahan ini, meskipun kecil, tetap penting untuk diperhatikan, terutama dalam aplikasi nyata, karena salah mengklasifikasikan email penting sebagai Spam dapat berdampak negatif.

Tabel 4. Tabel Evaluasi Model Pengujian Data Testing

	precision	recall	f1-score	support
0	0.98	0.99	0.99	541
1	0.98	0.96	0.97	274
accuracy			0.98	815
macro avg	0.98	0.98	0.98	815
weighted avg	0.98	0.98	0.98	815

Akurasi:	0.98
Presisi:	0.98
Recall:	0.98
F1-Score:	0.98

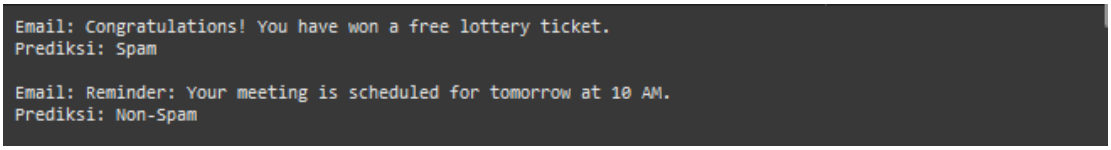
Berdasarkan tabel 4 diatas menunjukan Model klasifikasi email menunjukkan performa yang sangat baik dengan akurasi sebesar 98%, yang berarti sebagian besar prediksi model sesuai dengan label asli. Presisi sebesar 98% menunjukkan bahwa prediksi spam dan non-spam sebagian besar benar, sementara recall sebesar 98% mengindikasikan model mampu mendeteksi hampir semua email spam dan non-spam dalam dataset. Dengan F1-Score juga mencapai 98%, model membuktikan keseimbangan yang kuat antara presisi dan recall, sehingga konsisten dalam menangani kedua kategori secara efektif. Hasil ini menunjukkan bahwa model *Logistic Regression* sangat andal untuk mendeteksi email spam dan non-spam, dengan performa yang hampir sempurna dalam tugas klasifikasi ini.

Tabel 5. Pengujian Pada Data Testing

No	Teks Asli	Prediksi	Label Asli
1	uk ga data hi vinc got forward request deal uk ga data requir regard anjam x forward kyran hank lon ect margaret carson enron vinc…….	Bukan Spam	Bukan Spam
2	alp present behalf enron corp would like invit alp project present group student jess h jone graduat school manag rice univers student……	Bukan Spam	Bukan Spam
3	medicin differ price big save brand name drug injustic anywher threat justic everywher rise sin virtu fall judg condemn crimin absolv height path pave dagger	Spam	Spam
4	list major search engin submit websit search engin may increas onlin sale dramat invest time money websit simpli must submit websit oniine……	Spam	Spam
5	alp present vinc thank invit attend present anoth commit dinner pleas indic specif room present known thank wil wrote behalf enron corp would like invit alp project present group student jess h jone graduat school…….	Bukan Spam	Bukan Spam
6	download softwar http rosari realoemsal com	Spam	Spam
7	local softwar languag avail hello would like offer local softwar version german french spanish uk mani other aii list softwar	Spam	Spam

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Langkah selanjutnya dalam penelitian ini adalah menganalisis konten email baru dengan tujuan untuk menentukan apakah konten tersebut termasuk dalam kategori spam atau non-spam. Ini adalah langkah terakhir dari proses penelitian, dengan tujuan untuk mengevaluasi keefektifan metode yang digunakan. Hasil dari penelitian ini dapat dilihat lebih jelas pada gambar 4, yang memberikan representasi visual dari hasil klasifikasi untuk memberikan gambaran yang lebih jelas mengenai kinerja metode.



Gambar 4. Hasil Pengujian Dengan Menggunakan Konten Email Baru

Gambar 4, menunjukan Hasil pengujian model yang berhasil mengklasifikasikan email dengan baik. Email pertama, "Congratulations! You have won a free lottery ticket," diprediksi sebagai Spam, sesuai dengan ciri-ciri umum email spam seperti janji hadiah gratis. Sementara itu, email kedua, "Reminder: Your meeting is scheduled for tomorrow at 10 AM," diprediksi sebagai Non-Spam, menunujukan konten yang relevan dan formal. Prediksi yang akurat ini menunjukkan keandalan model dalam mengenali pola teks email baru.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma Logistic Regression untuk klasifikasi email spam dengan akurasi, presisi, recall, dan F1-Score masing-masing sebesar 98%. Tahap pra-pemrosesan data seperti case folding, tokenizing, dan stemming, serta ekstraksi fitur menggunakan TF-IDF, terbukti meningkatkan akurasi model. Logistic Regression mampu mengklasifikasikan email spam dan non-spam secara akurat dengan sedikit kesalahan, menunjukkan potensi penerapannya dalam sistem penyaringan email untuk meningkatkan efisiensi komunikasi dan mencegah risiko serangan phishing. Hasil ini menunjukkan peningkatan performa dibandingkan penelitian sebelumnya, yang hanya mencapai akurasi sebesar 97%. Temuan ini menegaskan bahwa Logistic Regression adalah metode yang andal untuk klasifikasi email spam.

DAFTAR PUSTAKA

[1] S. Ulya, M. A. Soeleman, and F. Budiman, “Optimasi Parameter K Pada Algoritma K-NN Untuk Klasifikasi Prioritas Bantuan Pembangunan Desa,” *Techno.Com*, vol. 20, no. 1, pp. 83–96, 2021, doi: 10.33633/tc.v20i1.4215.

[2] N. L. W. S. R. Ginantra *et al.*, *Data Mining dan Penerapan Algoritma*. 2021.

[3] H. Hendra, B. T. S. SP., and A. A. Setyawan, “Menggunakan Binary Classification Untuk Mendeteksi Spam Pada Sms Dengan Metode Logistic Regression,” *CONTEN Comput. Netw. Technol.*, vol. 4, no. 1, pp. 21–30, 2024, doi: 10.31294/conten.v4i1.3543.

[4] A. Nur, R. Hasanah, R. A. Krestianti, and S. Wati, “Implementasi Algoritma Regresi Logistik untuk Binary Classification dalam Spam SMS dan WhatsApp,” *Agustus*, vol. 7, pp. 2549–7952, 2023, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/inotex/

[5] M. A. I. Hutagalung and Sutarman, “Penalized Maximum Likelihood Estimation dengan Algoritma Gradient descent pada Model Regresi Logistik Multinomial,” *IJM Indones. J. Multidiscip.*, vol. 2, no. 6 SE-Articles, pp. 673–683, 2024, [Online]. Available: http://journal.csspublishing.com/index.php/ijm/article/view/951

[6] A. N. Salim, A. Adryani, and T. Sutabri, “Deteksi Email Spam dan Non-Spam Berdasarkan Isi Konten Menggunakan Metode K-Nearest Neighbor dan Support Vector Machine,” *Syntax Idea*, vol. 6, no. 2, pp. 991–1001, 2024, doi: 10.46799/syntax-idea.v6i2.3052.

[7] E. P. Laksono and A. Wicaksono, “Penyaringan Spam email menggunakan K-Means,” *J.*

Spektro, vol. 5, no. 2, pp. 26–32, 2022, [Online]. Available: https://ejurnal.undana.ac.id/index.php/spektro/article/view/9531%0Ahttps://ejurnal.undana.ac.id/index.php/spektro/article/download/9531/4687

[8] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, pp. 46–53, 2015.

[9] Brury Barth Tangkere, "Analisis Performa Logistic Regression dan Support Vector Classification untuk Klasifikasi Email Phising," *J. Ekon. Manaj. Sist. Inf.*, vol. 5, no. 4, pp. 442–450, 2024, doi: 10.31933/jemsi.v5i4.1916.

[10] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2023.

[11] E. Laksono, A. Basuki, and F. Bachtiar, "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 377–383, 2020, doi: 10.29207/resti.v4i2.1845.

[12] J. Junifer Pangaribuan, H. Tanjaya, and K. Kenichi, "Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression," *J. Inf. Syst. Dev.*, vol. 06, no. 02, pp. 1–10, 2021.

[13] A. S. Lukmanul Hakim, *Introduction to Machine Learning Using R*. IPB Press, 2022.

[14] I. Ir. Heliza Rahmania Hatta, S.Kom., M.Kom. *et al.*, *Intelligent Systems*, no. June. Kota Batam: Penerbit Yayasan Cendikia Mulia Mandiri, 2024.