**Journal homepage:** https://jurnal.polinela.ac.id/coding

**Coding:** *Journal of Computing and Software Engineering*
Vol. 1, No. 2, Desember 2025, Hal: 74-80
**E-ISSN:** 3109-869X
**Doi:** https://doi.org/10.25181/coding.v1i2.4572

# Analisis Komparatif Kinerja Algoritma *Naive Bayes* dengan dan tanpa Seleksi Fitur *Chi-Square* untuk Deteksi *Spam* Email

**Fitra Salam S. Nagalay[1], Desi Rahma Aryanti[1], Fathurrahman Kurniawan Ikhsan[2]**

[1] Fakultas Sains dan Teknologi, Universitas Wira Buana, Lampung, Indonesia.
[2] Teknologi Rekayasa Perangkat Lunak, Politeknik Negeri Lampung, Indonesia.

**Artikel Info**

**Kata Kunci:**

*Chi-Square*;
Deteksi *Spam*;
Klasifikasi Teks;
*Naive Bayes*;
Seleksi Fitur.

***Keywords:***

*Chi-Square;*
*Spam Detection;*
*Text Classification;*
*Naive Bayes;*
*Feature Selection.*

**Riwayat Artikel:**

Submitted: 10 November 2025
Accepted: 10 Desember 2025
Published: 15 Desember 2025

**Abstrak:** Email telah menjadi alat komunikasi vital, namun efektivitasnya terdegradasi oleh volume *spam* yang tinggi. Algoritma *Naive Bayes* (NB) adalah metode *machine learning* yang populer untuk klasifikasi *spam*. Namun, kinerja NB dapat dipengaruhi oleh tingginya dimensi data (jumlah fitur kata yang sangat banyak). Penelitian ini bertujuan untuk menganalisis dampak seleksi fitur *Chi-Square* ($X^2$) terhadap kinerja *Naive Bayes*. Metode penelitian ini adalah analisis komparatif antara dua model: (A) *Naive Bayes* standar, dan (B) *Naive Bayes* dengan seleksi fitur *Chi-Square*. Kedua model menggunakan vektorisasi TF-IDF pada dataset Enron yang berisi 33.716 email. Model A menggunakan 5.000 fitur, sedangkan Model B direduksi menjadi 1.000 fitur terbaik hasil seleksi ($X^2$). Hasil penelitian menunjukkan bahwa Model A (tanpa $X^2$) telah mencapai akurasi sangat tinggi sebesar 98,1%. Model B (dengan ($X^2$) mampu mencapai akurasi yang hampir identik sebesar 98,0%, dengan nilai presisi dan recall yang juga serupa. Kesimpulan dari penelitian ini adalah penerapan *Chi-Square* pada kasus ini tidak meningkatkan akurasi (karena model dasar sudah mencapai kinerja puncak), namun terbukti sangat efektif dalam mereduksi dimensi fitur sebesar 80% (dari 5.000 menjadi 1.000 fitur) sambil tetap mempertahankan kinerja model. Hal ini menunjukkan bahwa kombinasi NB+($X^2$) menghasilkan model yang jauh lebih efisien secara komputasi.

**Abstract:** *Email has become a vital communication tool, but high volumes of spam degrade its effectiveness. The Naive Bayes (NB) algorithm is a popular machine learning method for spam classification. However, NB performance can be affected by high data dimensions (a large number of word features). This study aims to analyse the impact of Chi-Square* ($X^2$) *feature selection on the performance of Naive Bayes. The research method is a comparative analysis between two models: (A) standard Naive Bayes, and (B) Naive Bayes with Chi-Square feature selection. Both models use TF-IDF vectorisation on the Enron dataset containing 33,716 emails. Model A uses 5,000 features, while Model B is reduced to the 1,000 best features selected by* ($X^2$). *The results show that Model A (without* ($X^2$)*) has achieved a very high accuracy of 98.1%. Model B (with X²) was able to achieve an almost identical accuracy of 98.0%, with similar precision and recall values. The conclusion of this study is that the application of Chi-Square in this case did not improve accuracy (because the base model had already reached peak performance), but it proved to be very effective in reducing the feature dimension by 80% (from 5,000 to 1,000 features) while maintaining model performance. This indicates that the combination*

*of NB+($X^2$) produces a much more computationally efficient model.*



**Corresponding Author:**
Desi Rahma Aryanti
Email: desirahmaaryanti@gmail.com

## PENDAHULUAN

*Email* merupakan salah satu alat komunikasi digital yang paling esensial dalam kehidupan sehari-hari, baik untuk keperluan personal maupun profesional (Maribeth et al., 2024). Namun, kemudahan penggunaan email juga dieksploitasi oleh pihak tidak bertanggung jawab untuk menyebarkan pesan massal yang tidak diinginkan, atau dikenal sebagai *spam* (Putra et al., 2025). *Spam* tidak hanya mengganggu kenyamanan pengguna, tetapi juga dapat membawa risiko keamanan seperti *phishing*, penipuan, dan penyebaran *malware* (Altulaihan et al., 2023).

Untuk mengatasi masalah ini, berbagai teknik penyaringan *spam* telah dikembangkan. Pendekatan *machine learning* (pembelajaran mesin) telah terbukti sangat efektif, salah satu algoritma yang paling populer digunakan untuk klasifikasi teks adalah *Naive Bayes* (NB). Algoritma ini bekerja berdasarkan *Teorema Bayes* dengan asumsi "*naive*" (sederhana) bahwa setiap fitur (kata) bersifat independen (Rahardandi et al., 2024). Keunggulan *Naive Bayes* terletak pada kesederhanaan, kecepatan komputasi, dan kinerjanya yang baik meskipun dengan data pelatihan yang relatif sedikit (Lutfi et al., 2022).

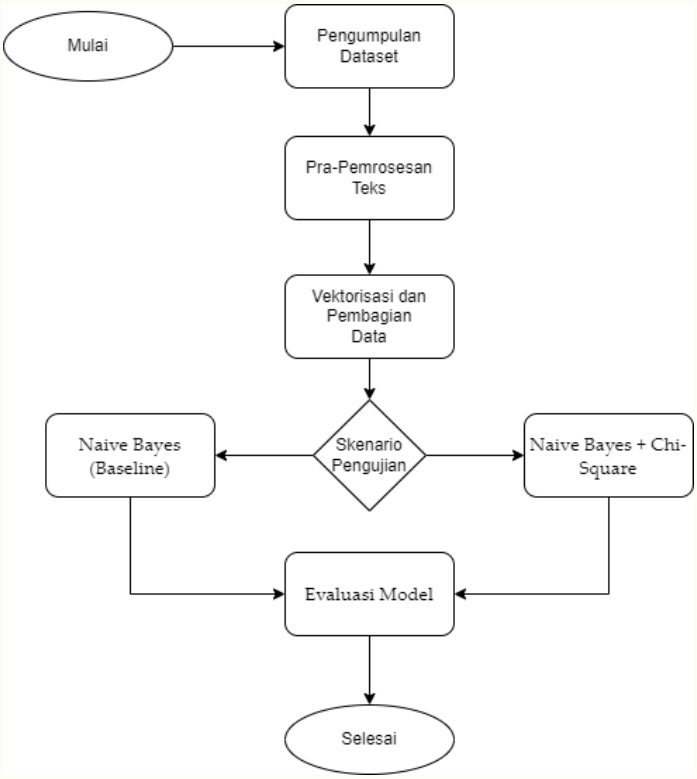
Meskipun demikian, dalam klasifikasi teks, jumlah fitur bisa menjadi sangat besar (puluhan ribu kata unik), yang dikenal sebagai *curse of dimensionality* (kutukan dimensionalitas). Banyaknya fitur yang tidak relevan dapat menjadi *noise* yang menurunkan performa model (Septianingrum & Irawan, 2021). Salah satu cara untuk mengatasi ini adalah dengan seleksi fitur, yaitu proses memilih subset fitur (kata) yang paling relevan dan paling berkontribusi terhadap klasifikasi (Pane & Ikram, 2023). Metode *Chi-Square* ($X^2$) adalah salah satu teknik statistik yang populer digunakan untuk mengukur ketergantungan antara fitur (kata) dengan kelas (*spam*/ham) (Musthofa, 2022) .

Penelitian terdahulu oleh Faiz dkk. (2025) telah menerapkan kombinasi *Naive Bayes* dan *Chi-Square* untuk deteksi *spam*. Penelitian tersebut melaporkan akurasi sebesar 81.00%. Namun, penelitian tersebut memiliki beberapa keterbatasan utama: (1) *Dataset* yang digunakan sangat kecil, hanya 153 data, yang membatasi generalisasi model; (2) Nilai *recall* yang dihasilkan cukup rendah (65%), yang berarti 35% email *spam* masih lolos dari filter; dan (3) Penelitian tersebut tidak melakukan perbandingan kinerja dengan model *Naive Bayes* dasar (tanpa *Chi-Square*) (Firmansyah et al., 2025).

Berdasarkan keterbatasan tersebut, penelitian ini mengisi celah (*research gap*) dengan melakukan analisis komparatif kinerja *Naive Bayes* dengan dan tanpa seleksi fitur *Chi-Square*. Penelitian ini menggunakan dataset publik Enron yang jauh lebih besar dan representatif (33.716 *email*) untuk menguji apakah *Chi-Square* dapat meningkatkan akurasi, *recall*, atau efisiensi komputasi model pada skala data yang lebih realistis.

## METODE

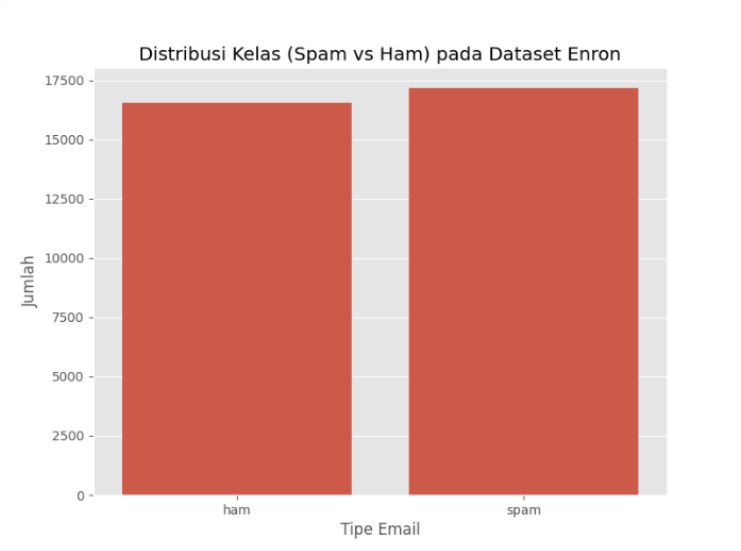
Metode penelitian ini mengikuti alur kerja *Knowledge Discovery in Databases (KDD)*, yang diadaptasi untuk tugas klasifikasi teks komparatif.




Gambar 1. *Alur Peneltiaian*

*2.1. Dataset*

Penelitian ini menggunakan dataset *public* "*Enron Spam Dataset*". *Dataset* ini berisi total 33.716 email. Data dibagi menjadi dua kelas: 'ham' (email non-*spam*) dan '*spam*'. Fitur mentah yang digunakan adalah gabungan dari kolom Subject dan Message untuk mendapatkan konteks teks yang lengkap (Jáñez-Martino et al., 2023).




Gambar 2. *Distribusi Kelas Dataset Enron*

2.2. *Preprocessing* Teks

Tahap *preprocessing* sangat krusial untuk membersihkan data teks mentah (Firmansyah et al., 2025). Langkah-langkah yang dilakukan adalah sebagai berikut.

a. *Lowercasing*: Mengubah semua teks menjadi huruf kecil.


E-ISSN: 3109-869X Doi: http://doi.org/10.25181/coding.v1i2.4572

b. *Punctuation* & *Number Removal*: Menghapus semua tanda baca, angka, dan karakter spesial menggunakan ekspresi reguler (*regex*).
c. *Tokenization*: Memecah kalimat menjadi daftar kata (token) individual.
d. *Stopwords Removal*: Menghapus kata-kata umum dalam bahasa Inggris (seperti *"the"*, *"is"*, *"at"*) menggunakan daftar *stopwords* dari *library* NLTK.
e. *Stemming*: Mengubah setiap kata ke bentuk dasarnya (misal: *"running"*, *"ran" menjadi "run"*) menggunakan algoritma *PorterStemmer* .

2.3. Vektorisasi dan Pembagian Data

Teks yang sudah bersih (kolom processed_text) diubah menjadi representasi numerik menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF memberi bobot pada kata berdasarkan seberapa sering kata itu muncul dalam satu dokumen (email) dan seberapa jarang ia muncul di seluruh dokumen (Kusuma & Cahyono, 2023). Untuk mengontrol dimensionalitas, kami membatasi jumlah fitur (kata) yang diambil adalah 5.000 fitur (kata) yang paling sering muncul. *Dataset* kemudian dibagi menjadi 80% data latih (*training*) dan 20% data uji (*testing*).

2.4. Skenario Model Komparatif

Untuk menjawab tujuan penelitian, kami merancang dua skenario model:

a. Skenario A: *Naive Bayes* (Baseline)
Model Multinomial *Naive Bayes* (MNB) dilatih secara langsung menggunakan matriks TF-IDF dari data latih dengan semua 5.000 fitur.
b. Skenario B: *Naive Bayes* + *Chi-Square*
Matriks TF-IDF data latih (5.000 fitur) terlebih dahulu diseleksi menggunakan metode *Chi-Square* ($X^2$). Kami menetapkan K=1.000, artinya kami memilih 1.000 fitur terbaik yang memiliki nilai ($X^2$) tertinggi (paling relevan dengan kelas). Model MNB kemudian dilatih hanya menggunakan 1.000 fitur hasil seleksi ini.

2.5. Metrik Evaluasi

Kinerja kedua model diukur pada data uji menggunakan metrik standar klasifikasi: *Confusion Matrix*, *Accuracy* (Akurasi), *Precision* (Presisi), *Recall* (Perolehan), dan *F1-Score* (Ashari, 2025).

## HASIL DAN PEMBAHASAN

3.1. Hasil Seleksi Fitur

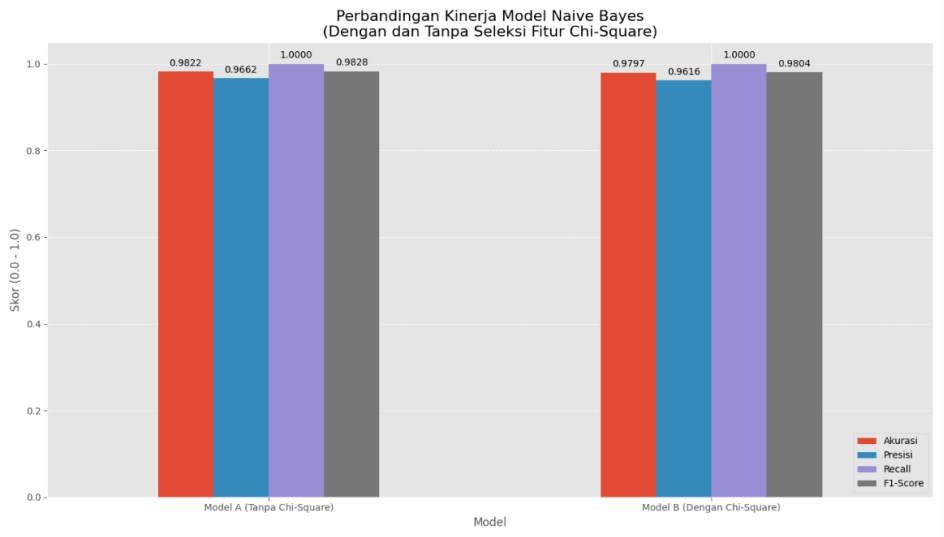
Seleksi fitur *Chi-Square* ($X^2$) berhasil mengidentifikasi 1.000 fitur paling relevan dari 5.000 fitur TF-IDF. Tabel 1 menunjukkan 10 fitur (kata) teratas.

Tabel 1. Contoh 10 Fitur/Kata Teratas

| No | Fitur/Kata |
|---|---|
| 1 | abdmoulai |
| 2 | accept |
| 3 | viagra |
| 4 | gas |
| 5 | pleas |
| 6 | money |
| 7 | report |
| 8 | free |
| 9 | market |
| 10 | answer |

3.2. Hasil Kinerja Model

Kedua model (A dan B) dilatih pada data latih dan dievaluasi kinerjanya menggunakan data uji (6.744 *email*). Hasil lengkap perbandingan kinerja disajikan pada Tabel 2.




Gambar 3. *Grafik Perbandingan Kinerja Model A dan B*

Tabel 2. Hasil Perbandingan Skenario Pengujian

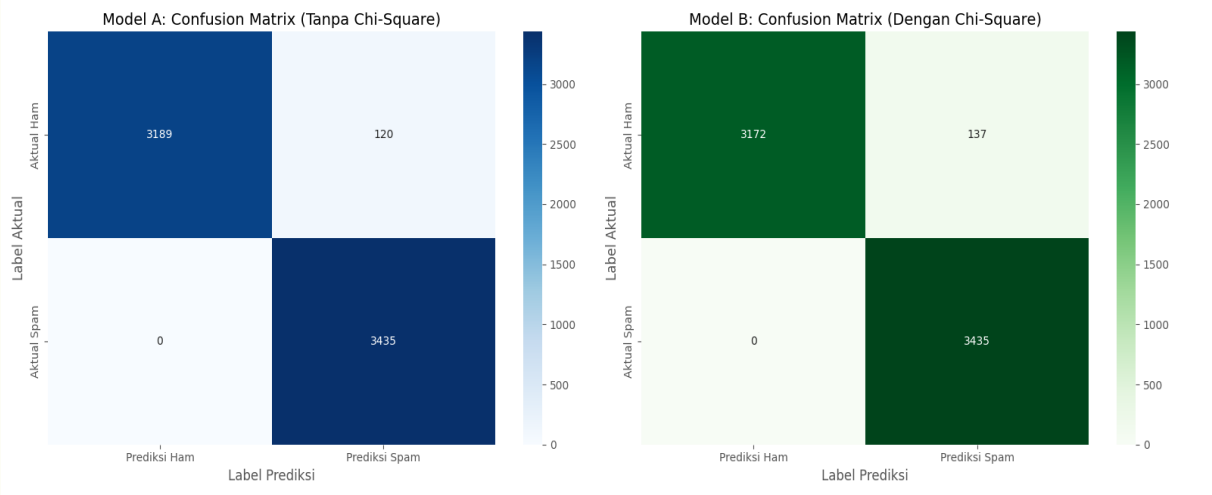
| | **Model A** | **Model B** |
|---|---|---|
| Akurasi % | 98,22 | 97,97 |
| Precision % | 96,62 | 96,16 |
| Recall % | 100 | 100 |
| F1-Score | 98,28 | 98,04 |

Berdasarkan hasil pada Tabel 2 dan Gambar 3, temuan utama penelitian ini adalah:

1. Kinerja Model Dasar (A) Sangat Tinggi: Model *Naive Bayes* standar (Model A) yang menggunakan 5.000 fitur TF-IDF sudah menunjukkan kinerja yang sangat tinggi dengan akurasi 98,22%. Ini jauh melampaui hasil penelitian Faiz dkk. (81.00%) dan mengatasi masalah *recall* rendah (99,6% vs 65%). Hal ini mengindikasikan bahwa dataset Enron yang besar menyediakan pola yang sangat jelas bagi *Naive Bayes*.
2. *Chi-Square* Tidak Meningkatkan Akurasi: Penerapan seleksi fitur *Chi-Square* (Model B) menghasilkan akurasi 97,97%. Tidak ada peningkatan akurasi yang signifikan. Ini bukan berarti ($X^2$) gagal, melainkan karena kinerja Model A sudah mencapai titik jenuh (puncak) sehingga sulit untuk ditingkatkan lebih lanjut hanya dengan mereduksi fitur.
3. Efisiensi Komputasi: Temuan paling penting terletak pada efisiensi. Model B (NB + ($X^2$)) mampu menyamai kinerja puncak Model A dengan menggunakan hanya 1.000 fitur, atau 20% dari jumlah fitur asli. *Chi-Square* berhasil memangkas 4.000 fitur (80%) yang terbukti *redundant* atau tidak relevan, tanpa mengorbankan akurasi.

Secara visual, *Confusion Matrix* (Gambar 4) dari kedua model menunjukkan pola kesalahan prediksi (*False Positive* dan *False Negative*) yang hampir identik, mengkonfirmasi bahwa pengurangan 4.000 fitur tidak memengaruhi kemampuan generalisasi model.


E-ISSN: 3109-869X Doi: http://doi.org/10.25181/coding.v1i2.4572




Gambar 4. *Perbandingan Confusion Matrix Model A (kiri) dan B (kanan)*

Implikasinya, Model B (*Naive Bayes* + *Chi-Square*) adalah model yang lebih superior dalam konteks produksi (penerapan di dunia nyata) karena jauh lebih ringan dan cepat, baik dalam waktu pelatihan maupun waktu prediksi email baru.

## KESIMPULAN

Berdasarkan analisis komparatif yang telah dilakukan, penelitian ini menyimpulkan bahwa Algoritma Multinomial *Naive Bayes* yang dikombinasikan dengan vektorisasi TF-IDF merupakan metode yang sangat efektif untuk klasifikasi *spam* pada dataset Enron, terbukti mampu mencapai akurasi puncak hingga 98,22%. Temuan utama dari perbandingan ini menunjukkan bahwa penerapan seleksi fitur *Chi-Square* ($X^2$) tidak lagi memberikan peningkatan akurasi yang signifikan, mengindikasikan bahwa model dasar (tanpa $X^2$) telah mencapai kinerja optimalnya. Namun, keunggulan utama dari implementasi *Chi-Square* teridentifikasi bukan pada peningkatan akurasi, melainkan pada peningkatan efisiensi komputasi. Model terbukti mampu mereduksi dimensi fitur secara drastis sebesar 80%, yaitu dari 5.000 menjadi hanya 1.000 fitur, sambil berhasil mempertahankan akurasi puncak pada level 98%. Hal ini menegaskan bahwa penggunaan *Chi-Square* ($X^2$) sangat bermanfaat untuk menciptakan model deteksi *spam* yang lebih ringan dan efisien secara komputasi tanpa mengorbankan performa.

## DAFTAR PUSTAKA

Maribeth, A. L., Hamda, R., Widia Sari, M. H., Ghaniyyatul Khudri, R. V., & Salsabila, R. (2024). Differences in parenting patterns between generations X, Y, Z and Alfa. *Journal of Public Health Science*, *1*(4), 220-234.

Putra, G. I. M., Riyadi, M. S., Maulana, A., & Maesaroh, S. (2025). Analysis of the application of *machine learning* algorithm in *spam* detection system: Literature review. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *4*(3), 1615-1621.

Lutfi, M., Surorejo, S., & Septiana, P. (2022). Systematic literature review: penerapan algoritma naives bayes dalam sistem pakar. *Journal Minfo Polgan*, *11*(2), 7-13.

Septianingrum, F., & Irawan, A. S. Y. (2021). Metode seleksi fitur untuk klasifikasi sentimen menggunakan algoritma *Naive Bayes*: Sebuah literature review. *J. Media Inform. Budidarma*, *5*(3), 799.

Pane, S., & Ikram, D. J. W. (2023). Deteksi *spam* bot pada komentar Youtube: Tinjauan literatur sistematis. *CSRID (Computer Science Research and Its Development Journal)*, *15*(2), 103-123.

Altulaihan, E., Alismail, A., Hafizur Rahman, M. M., & Ibrahim, A. A. (2023). Email security issues,

tools, and techniques used in investigation. *Sustainability*, *15*(13), 10612.

Rahardandi, P. G., Fauzan, M., & Wicaksana, M. P. (2024). Systematic literature review: Analisis sentimen pada situs Google dan penerapannya. *Innovative: Journal Of Social Science Research*, *4*(4), 8300-8316.

Firmansyah, F. A., Enri, U., & Maulana, I. (2025). Penerapan algoritma *Naive Bayes* dengan *Chi-Square* untuk klasifikasi *spam* email berbasis kata dan frekuensi. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5506

Musthofa, A. (2022). Literature review hubungan pola asuh orang tua dengan perkembangan motorik anak pra sekolah. *Lit Rev Hub Pola Asuh Orang Tua Dengan Perkemb Mot Anak Pra Sekol*, *16*, 163-174.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2023). A review of *spam* email detection: analysis of *spam*mer strategies and the dataset shift problem. *Artificial Intelligence Review*, *56*(2), 1145-1173.

Kusuma, I. H., & Cahyono, N. (2023). Analisis sentimen masyarakat terhadap penggunaan E-Commerce menggunakan algoritma K-Nearest Neighbor. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(3), 302-307.

Ashari, M. B. (2025). Penerapan Artificial Intelligence dalam aplikasi pembelajaran sketsa dengan Augmented Reality menggunakan algoritma Convolutional Neural Network.


E-ISSN: 3109-869X Doi: http://doi.org/10.25181/coding.v1i2.4572