Evaluasi Kinerja Model Klasifikasi pada Sistem Deteksi Spam Email Menggunakan Confusion Matrix

1th Siti Rahma Alia
20230040023
Teknik Informatika
Universitas Nusa Putra Sukabumi
siti.rahma_ti23@nusaputra.ac.id

Abstrak— Spam email merupakan salah satu permasalahan utama dalam sistem komunikasi digital karena dapat mengganggu pengguna serta berpotensi membawa ancaman keamanan. Berbagai pendekatan machine learning telah dikembangkan untuk mendeteksi spam email secara otomatis, namun diperlukan evaluasi kinerja model agar sistem yang dibangun dapat bekerja secara optimal. Penelitian ini bertujuan untuk mengevaluasi kinerja model klasifikasi dalam mendeteksi spam email menggunakan confusion matrix dan metrik evaluasi klasifikasi. Dataset yang digunakan adalah dataset email berbahasa Indonesia yang terdiri dari dua kelas, yaitu spam dan non-spam. Model klasifikasi yang digunakan dalam penelitian ini adalah Naive Bayes dengan representasi fitur TF-IDF. Evaluasi dilakukan menggunakan confusion matrix serta metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model mampu mengklasifikasikan email spam dan non-spam dengan tingkat akurasi yang tinggi. Meskipun demikian, masih ditemukan sejumlah kesalahan klasifikasi yang perlu dianalisis lebih lanjut. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem deteksi spam email berbasis machine learning.

Kata Kunci— (deteksi spam email, klasifikasi, Naive Bayes, confusion matrix, evaluasi model)

I. Latar Belakang

Perkembangan teknologi informasi menyebabkan penggunaan email sebagai media komunikasi digital semakin meningkat. Email digunakan secara luas dalam berbagai aktivitas, baik untuk keperluan pribadi, akademik, maupun bisnis. Namun, peningkatan penggunaan email tersebut juga diiringi dengan maraknya penyebaran spam email yang berisi iklan tidak relevan, penipuan, serta konten berbahaya. Keberadaan spam email tidak hanya mengganggu kenyamanan pengguna, tetapi juga berpotensi menimbulkan ancaman keamanan seperti pencurian data dan penyebaran malware.

Berbagai penelitian telah dilakukan untuk mengatasi permasalahan spam email dengan memanfaatkan teknik machine learning. Model klasifikasi seperti Naive Bayes, Support Vector Machine, Random Forest, hingga pendekatan deep learning telah digunakan untuk membedakan email spam dan non-spam [2][3][6]. Selain pemilihan algoritma yang tepat, evaluasi kinerja model menjadi aspek penting untuk memastikan bahwa sistem deteksi spam yang dibangun dapat bekerja secara efektif dan andal.

Evaluasi kinerja model klasifikasi umumnya dilakukan menggunakan confusion matrix dan metrik turunannya, seperti accuracy, precision, recall, dan F1-score. Confusion matrix memberikan gambaran yang jelas mengenai hasil prediksi model serta jenis kesalahan klasifikasi yang terjadi, termasuk false positive dan false negative. Oleh karena itu, penelitian ini berfokus pada evaluasi kinerja model klasifikasi dalam mendeteksi spam email menggunakan confusion matrix sebagai alat evaluasi utama.

A. Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kinerja model klasifikasi dalam mendeteksi spam email menggunakan dataset email berbahasa Indonesia?
2. Bagaimana hasil evaluasi kinerja model berdasarkan confusion matrix dan metrik evaluasi klasifikasi?
3. Kesalahan klasifikasi apa saja yang paling sering terjadi pada sistem deteksi spam email?

B. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Melatih model klasifikasi untuk mendeteksi spam email menggunakan dataset berbahasa Indonesia.
2. Mengevaluasi kinerja model klasifikasi menggunakan confusion matrix dan metrik evaluasi.
3. Menganalisis kesalahan klasifikasi yang terjadi pada hasil prediksi model.

C. Batasan Masalah

Penelitian ini dibatasi oleh:

1. Dataset yang digunakan merupakan dataset email berbahasa Indonesia.
2. Klasifikasi hanya terdiri dari dua kelas, yaitu spam dan non-spam.
3. Model yang digunakan adalah Naive Bayes sebagai model klasifikasi.

4. Evaluasi kinerja model dibatasi pada confusion matrix dan metrik evaluasi klasifikasi.
5. Penelitian tidak membahas implementasi sistem secara real-time.

II. Landasan Teori

A. Deteksi Spam Email

Deteksi spam email merupakan proses klasifikasi email ke dalam kategori spam atau non-spam berdasarkan konten dan karakteristik tertentu. Pendekatan berbasis machine learning memungkinkan sistem untuk belajar dari data historis dan mengenali pola yang membedakan email spam dan non-spam [2].

B. Algoritma Naive Bayes

Naive Bayes merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi antar fitur. Algoritma ini banyak digunakan dalam klasifikasi teks, termasuk deteksi spam email, karena sederhana, efisien, dan mampu memberikan performa yang baik pada dataset teks berdimensi tinggi [2][5].

C. Confusion Matrix dan Matrik Evaluasi

Confusion matrix adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi dengan membandingkan label aktual dan hasil prediksi. Dari confusion matrix dapat dihitung berbagai metrik evaluasi seperti accuracy, precision, recall, dan F1-score. Metrik ini sering digunakan dalam penelitian deteksi spam email untuk menilai efektivitas model klasifikasi [6].

III. Metodologi Penelitian

Penelitian ini dilakukan menggunakan Google Colab dengan bahasa pemrograman Python dan library pendukung seperti pandas, scikit-learn, matplotlib, dan seaborn.

A. Dataset

Dataset yang digunakan merupakan dataset email berbahasa Indonesia yang terdiri dari 2.636 data email dengan dua kelas, yaitu spam dan non-spam. Dari dataset tersebut dipilih sebanyak 1.000 data untuk digunakan dalam penelitian ini.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2636 entries, 0 to 2635
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Kategori  2636 non-null   object
 1   Pesan     2636 non-null   object
dtypes: object(2)
memory usage: 41.3+ KB
```

Kategori	count
spam	1368
ham	1268

Gambar 1. Contoh dataset email spam dan non-spam

B. Pembagian Dataset

Dataset dibagi menjadi data latih dan data uji menggunakan metode train-test split dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Pembagian ini bertujuan untuk menguji kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

C. Ekstraksi Fitur

Ekstraksi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) untuk mengubah data teks email menjadi representasi numerik yang dapat diproses oleh model klasifikasi.

D. Pelatihan Model

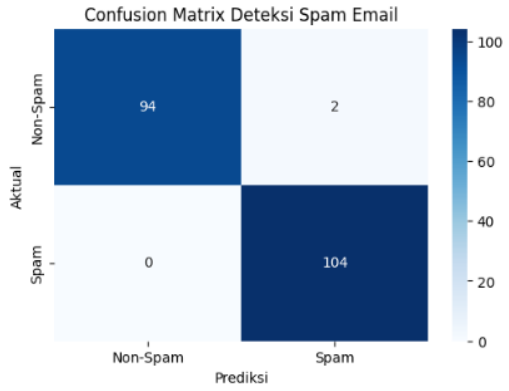
Model klasifikasi yang digunakan adalah Multinomial Naive Bayes. Model dilatih menggunakan data latih yang telah diekstraksi fiturnya, kemudian digunakan untuk memprediksi label pada data uji.

E. Evaluasi Model

Evaluasi model dilakukan menggunakan confusion matrix serta metrik evaluasi klasifikasi, yaitu accuracy, precision, recall, dan F1-score.

IV. Hasil dan Pembahasan

A. Hasil Confusion Matrix



Gambar 2. Confusion Matrix Deteksi Spam Email

Hasil confusion matrix menunjukkan bahwa sebagian besar email berhasil diklasifikasikan dengan benar oleh model. Nilai true positive dan true negative mendominasi, sementara jumlah false positive dan false negative relatif kecil.

B. Hasil Matrix Evaluasi

Berdasarkan hasil pengujian, model memperoleh nilai accuracy sebesar 99%. Nilai precision, recall, dan F1-score untuk kedua kelas juga menunjukkan hasil yang tinggi. Hal ini mengindikasikan bahwa model Naive Bayes mampu mendeteksi spam email dengan baik pada dataset yang digunakan.

C. Analisi kesalahan klasifikasi

Meskipun performa model tergolong tinggi, masih terdapat beberapa kesalahan klasifikasi berupa false positive

dan false negative. Kesalahan ini dapat disebabkan oleh kemiripan konten antara email spam dan non-spam, serta keterbatasan fitur yang digunakan.

V. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa model klasifikasi Naive Bayes mampu memberikan performa yang sangat baik dalam mendeteksi spam email menggunakan dataset berbahasa Indonesia. Evaluasi menggunakan confusion matrix dan metrik evaluasi klasifikasi menunjukkan tingkat akurasi yang tinggi dengan jumlah kesalahan klasifikasi yang relatif kecil.

Meskipun demikian, penelitian ini masih memiliki keterbatasan, terutama pada penggunaan satu model klasifikasi dan fitur teks yang sederhana. Penelitian selanjutnya dapat mengembangkan sistem dengan membandingkan beberapa algoritma klasifikasi atau menggunakan pendekatan deep learning untuk meningkatkan kinerja deteksi spam email.

REFERENCES

[1] M. B. M. Amin et al., "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model BERT."

[2] J. Al Amien, H. Mukhtar, dan M. A. Rucyat, "Filtering Spam Email Menggunakan Algoritma Naive Bayes."

[3] M. Rustam, A. Brotokuncoro, dan R. Roestam, "Deteksi Email Spam dengan Continuous Bag-of-Words dan Random Forest."

[4] C. M. Bachri dan W. Gunawan, "Deteksi Email Spam Menggunakan Algoritma Convolutional Neural Network (CNN)."

[5] B. Aditya, M. K. Wijaya, dan A. Prabowo, "Pendekatan Naive Bayes Campuran untuk Klasifikasi Email Spam."

[6] E. S. Ainun, U. Inayah, dan M. Ilmih, "Klasifikasi Email Spam dan Ham Menggunakan SVM, Naive Bayes, dan Logistic Regression."