UJIAN AKHIR SEMESTER

KECERDASAN BUATAN

Disusun Untuk Memenuhi Ujian Akhir Semester Mata Kuliah Kecerdasan Buatan

KLASIFIKASI SPAM SMS MENGGUNAKAN K-NEAREST NEIGHBORS DAN REPRESENTASI TF-IDF BERBASIS FITUR TEKS

Dosen Pengampu : Leni Fitriani, ST. M.Kom

Disusun Oleh:

Sandra Oktavia Krisdiana Putri

2306037

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN ILMU KOMPUTER

INSTITUT TEKNOLOGI GARUT

2025

Daftar isi :

1. Business Understanding

1.1 Latar Belakang Masalah

Dalam beberapa tahun terakhir, pesan singkat (SMS) masih menjadi salah satu media komunikasi yang banyak digunakan, baik untuk keperluan pribadi maupun bisnis. Namun, kemudahan dalam pengiriman pesan ini juga dimanfaatkan oleh pihak-pihak yang tidak bertanggung jawab untuk menyebarkan spam SMS, yaitu pesan massal yang tidak diinginkan dan sering kali bersifat merugikan. Spam tidak hanya mengganggu kenyamanan pengguna, tetapi juga berpotensi mengandung tautan phishing**,** penipuan finansial**,** maupun malware yang dapat mengancam privasi serta keamanan data pribadi(Reviantika, 2021).

Metode penyaringan spam secara tradisional, seperti pemblokiran berbasis kata kunci, semakin kurang efektif. Hal ini disebabkan karena para spammer terus mengembangkan teknik manipulasi baru, seperti penggunaan singkatan, variasi kata, serta struktur kalimat yang dimodifikasi untuk menghindari deteksi. Selain itu, volume spam yang sangat besar menjadikan penyaringan secara manual tidak mungkin dilakukan secara efisien (Apriansyah et al., 2024).

Oleh karena itu, dibutuhkan sebuah solusi yang lebih cerdas, otomatis, dan adaptif, yang mampu mengidentifikasi pesan spam secara akurat tanpa mengganggu pengiriman pesan sah (legit/ham). Salah satu pendekatan yang menjanjikan adalah penerapan algoritma machine learning, yang memungkinkan sistem untuk belajar dari data historis dan mengenali pola-pola yang khas pada pesan spam secara otomatis(Pramakrisna et al., 2022).

1.2 Tujuan Proyek

Tujuan utama dari proyek ini adalah untuk mengembangkan sistem klasifikasi pesan singkat (SMS) menggunakan algoritma machine learning yang mampu membedakan antara pesan spam dan ham secara otomatis. Tujuan khusus dari proyek ini antara lain:

- Membangun model klasifikasi berbasis K-Nearest Neighbors (KNN) yang dapat mengidentifikasi pesan spam berdasarkan representasi teks menggunakan TF-IDF dan fitur numerik seperti panjang pesan, jumlah kata, rasio huruf kapital, dan jumlah tanda seru.
- Mengevaluasi performa model berdasarkan metrik akurasi, precision, recall, dan F1-score, guna mengetahui efektivitas model dalam menangani data tidak seimbang.
- Mengurangi false positive, yaitu kasus di mana pesan sah (ham) salah diklasifikasikan sebagai spam, serta false negative, yaitu spam yang tidak berhasil dideteksi.
- Memberikan solusi sederhana namun dapat diimplementasikan dalam sistem penyaringan pesan otomatis, terutama untuk dataset kecil atau menengah, sebagai baseline model dalam pengembangan sistem deteksi spam berbasis teks.

1.3 Pengguna dan manfaat

Proyek ini ditujukan bagi berbagai kalangan pengguna dan stakeholder yang relevan dengan isu deteksi spam SMS, di antaranya:

a. Pengguna Umum

Individu yang menggunakan SMS sebagai media komunikasi, yang dapat terbantu dengan sistem ini untuk menghindari pesan spam, penipuan, dan tautan berbahaya.

b. Penyedia Layanan Telekomunikasi

Operator seluler seperti Telkomsel, Indosat, dan XL dapat memanfaatkan model ini sebagai sistem penyaringan awal berbasis pembelajaran mesin untuk membantu mendeteksi dan mengurangi spam sebelum diteruskan ke pengguna akhir.

c. Pelaku Bisnis dan Layanan Komersial

Perusahaan yang mengirim pesan otomatis kepada pelanggan (misalnya, notifikasi transaksi, OTP, atau promosi) dapat memanfaatkan model ini untuk memastikan pesan sah mereka tidak terfilter sebagai spam, serta untuk memfilter pesan masuk dalam sistem layanan pelanggan.

d. Pengembang Aplikasi dan Sistem SMS Gateway

Startup atau developer yang membangun aplikasi berbasis perpesanan (SMS gateway, chat bot, OTP services) dapat mengintegrasikan model ini sebagai fitur anti-spam awal yang ringan, cepat, dan tidak memerlukan pelatihan berulang.

e. Peneliti dan Komunitas Akademik

Mahasiswa, dosen, dan peneliti yang tertarik dalam bidang Natural Language Processing (NLP) dan klasifikasi teks dapat menggunakan proyek ini sebagai kasus dasar (baseline) untuk mengembangkan sistem klasifikasi spam yang lebih canggih, termasuk perbandingan dengan algoritma lain seperti Naive Bayes atau SVM.

2. Data Understanding

2.1 Sumber Data :

Dataset yang digunakan dalam penelitian ini diperoleh dari situs Kaggle, dengan judul *"SMS Spam Collection Dataset"* yang dikembangkan oleh Almeida et al. Dataset ini terdiri dari 5572 pesan SMS, yang telah dilabeli secara manual sebagai ham (pesan normal) dan spam (pesan tidak diinginkan).

Dataset tersedia dalam format CSV, sehingga dapat langsung digunakan dalam proses analisis dan pelatihan model machine learning. Dataset ini bersifat publik dan dapat diakses melalui tautan berikut:

Link : https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

2.2 Deskripsi Fitur :

a. Sampel tampilan 5 baris pertama dataset

Lima baris pertama yang ditampilkan menunjukkan dua kolom utama, yaitu:

- Kolom Category menyatakan label klasifikasi berupa ham (pesan normal) atau spam (pesan tidak diinginkan).
- Kolom Message berisi teks asli dari isi pesan SMS

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Gambar 1. 5 Baris Pertama Dataset

Gambar 1. menunjukkan bahwa sebagian besar pesan berisi komunikasi informal, dengan struktur kalimat bebas yang mencerminkan bahasa sehari-hari. Hal ini menjadi tantangan dalam pemrosesan karena bahasa dalam SMS sering kali tidak mengikuti kaidah baku, mengandung singkatan, dan campuran antara huruf dan simbol.

b. Deskripsi statistic fitur

```
Deskripsi Statistik Fitur:
       Category                 Message
count      5572                    5572
unique        2                    5157
top         ham  Sorry, I'll call later
freq       4825                      30
```

Gambar 2. Deskripsi statistic fitur

Gambar 2. menunjukkan bahwa format pesan sangat informal dan penuh variasi—menggunakan singkatan, huruf kapital berlebih, tanda baca berulang, dan struktur kalimat tidak baku. Hal ini merupakan tantangan khas dalam pemrosesan bahasa alami (*Natural Language Processing*).

2.3 Ukuran dan Format Data

```
Ukuran dataset (baris, kolom): (5572, 2)
```

Gambar 3. Ukuran Dataset

Dalam tahap ini, belum dilakukan penghapusan duplikat ataupun pembersihan lanjutan, karena fokus utama masih pada pemahaman struktur awal data.

2.4 Tipe Data dan Target Klasifikasi

```
Tipe Data Per Kolom:
Category    object
Message     object
dtype: object
```

Gambar 4. Tipe Data

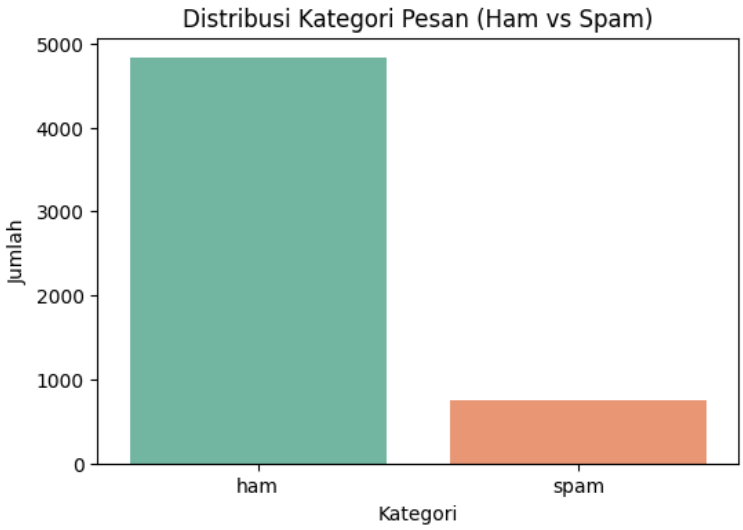
```
Distribusi Target Klasifikasi:
Category
ham     4825
spam     747
Name: count, dtype: int64


Tipe Data Target: object
```

Gambar 5. Distribusi Target Klasifikasi

3. Exploratory Data Analysis (EDA)

3.1 Visualisasi dan distribusi data

Distribusi kelas dalam dataset divisualisasikan menggunakan diagram batang (countplot), yang menunjukkan bahwa jumlah pesan ham jauh lebih banyak dibandingkan dengan spam. Hal ini merupakan indikasi awal bahwa dataset memiliki ketidakseimbangan kelas yang cukup signifikan.



Gambar 6. Distribusi Kategori Pesan

3.2 Analisis korelasi antar fitur

Sebelum dilakukan analisis korelasi, terlebih dahulu dilakukan proses feature engineering untuk mengekstrak informasi numerik dari pesan teks. Fitur-fitur yang ditambahkan antara lain:

- Message_Length: karena spam cenderung panjang (berisi ajakan, tautan, promosi).
- Word_Count: digunakan untuk mengukur jumlah informasi per pesan.

- Upper_Case_Ratio: karena huruf kapital sering digunakan untuk menonjolkan kata penting dalam spam.
- Exclamation_Count: spam sering kali menggunakan tanda seru secara berlebihan.

Setelah fitur numerik ditambahkan, dilakukan pemetaan label numerik (ham = 0, spam = 1). Korelasi antar fitur numerik divisualisasikan menggunakan heatmap. Hasilnya menunjukkan bahwa terdapat korelasi positif antara panjang pesan dengan jumlah kata, serta korelasi sedang antara rasio huruf kapital dan jumlah tanda seru dengan label spam.

Gambar 7. Korelasi Fitur

3.3 Deteksi data tidak seimbang

Analisis distribusi menunjukkan bahwa data termasuk imbalanced. Dari total 5157 pesan:

- Sekitar 87,6% merupakan pesan ham
- Hanya 12,4% yang merupakan pesan spam

```
Label
0    0.876
1    0.124
```

Gambar 8. Data Tidak Seimbang

Ketidakseimbangan ini perlu diperhatikan karena dapat memengaruhi performa model. Model klasifikasi dapat cenderung memprediksi kelas mayoritas jika tidak ditangani dengan tepat.

3.4 Insight awal dari pola data

Berdasarkan hasil eksplorasi data, mayoritas pesan dalam dataset merupakan pesan non-spam (ham) dengan proporsi sekitar 86.6%, sedangkan spam hanya sekitar 13.4%, menunjukkan adanya ketidakseimbangan kelas yang signifikan.

Selain itu, pesan spam cenderung memiliki jumlah tanda seru yang lebih tinggi dan rasio huruf kapital yang lebih besar, yang menunjukkan adanya pola penekanan atau urgensi. Korelasi positif antara panjang pesan dan jumlah kata juga logis karena semakin panjang suatu pesan maka umumnya semakin banyak kata yang terkandung di dalamnya (Wara Putera & Dewi Lestari, 2023).

Temuan ini dapat dimanfaatkan untuk mengekstrak fitur yang lebih informatif guna meningkatkan performa model klasifikasi spam.

4. Data Preparation

4.1 Pembersihan data

Langkah awal dalam tahap persiapan data adalah memastikan bahwa dataset dalam kondisi bersih dan siap untuk digunakan dalam proses pelatihan model. Pembersihan data dilakukan dengan

- Mengisi nilai kosong pada kolom Message dengan string kosong ('') agar tidak terjadi error saat proses ekstraksi fitur.
- Tidak dilakukan penghapusan duplikat pada tahap ini, sesuai keputusan eksplisit bahwa analisis awal dilakukan pada struktur data asli tanpa perubahan. Hal ini dipertahankan untuk menjaga jumlah data agar tetap utuh dan memungkinkan evaluasi model terhadap data sebagaimana adanya.

```
Jumlah data sebelum menghapus duplikat: 5572
Jumlah data setelah menghapus duplikat: 5157
Jumlah data training: 4125
Jumlah data testing : 1032
```

Gambar 9. Pembersihan Dataset

Pada Gambar 9. Setelah dilakukan penghapusan data duplikat, dataset berjumlah 5157 baris. Langkah ini penting untuk mencegah bias dan pengaruh data ganda dalam pelatihan model

4.2 Encoding target

```
# 3. Encoding target (sudah dilakukan, pastikan tetap ada)
data['Label'] = data['Category'].map({'ham': 0, 'spam': 1})
```

Gambar 10. Encoding Target

Karena algoritma machine learning tidak dapat langsung memproses data kategorikal, kolom Category perlu diubah ke dalam format numerik menggunakan teknik label encoding.

Hasil encoding disimpan dalam kolom baru bernama Label, yang menjadi target klasifikasi pada tahap pelatihan model

4.3 Standardisasi data

Fitur Message yang berupa teks tidak dapat digunakan langsung sebagai input numerik. Oleh karena itu, dilakukan proses ekstraksi fitur teks menggunakan TF-IDF

(Term Frequency–Inverse Document Frequency). TF-IDF akan menghasilkan representasi numerik untuk setiap pesan, berdasarkan seberapa penting kata-kata tertentu dalam pesan tersebut relatif terhadap seluruh korpus.

Karena nilai TF-IDF secara default sudah berada dalam rentang 0–1, maka proses standardisasi atau normalisasi tambahan tidak diperlukan.

4.4 Split data

Untuk mengevaluasi kinerja model secara objektif, dataset dibagi menjadi dua bagian:

- Training set (80%) digunakan untuk melatih model
- Testing set (20%) digunakan untuk menguji performa model pada data yang belum pernah dilihat sebelumnya

Pembagian ini menggunakan fungsi train_test_split dari scikit-learn, dengan parameter stratify=y untuk memastikan distribusi label spam dan ham tetap seimbang di kedua subset.

```
Distribusi label training:
 Label
0    0.876
1    0.124
Name: proportion, dtype: float64
Distribusi label testing:
 Label
0    0.876
1    0.124
Name: proportion, dtype: float64
```

Gambar 11. Split Data

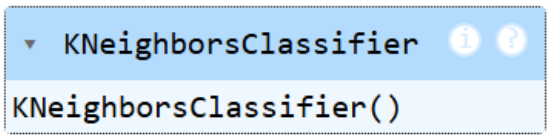
5. Modeling

Pada penelitian ini digunakan algoritma K-Nearest Neighbors (KNN) sebagai metode klasifikasi untuk membedakan antara pesan spam dan non-spam (ham). KNN merupakan salah satu algoritma non-parametrik yang sederhana namun cukup efektif dalam menyelesaikan masalah klasifikasi, termasuk untuk data berbasis teks yang telah direpresentasikan dalam bentuk numerik seperti TF-IDF (Arisona et al., 2023).

Alasan utama pemilihan algoritma KNN adalah sebagai berikut:

- Kesederhanaan konsep: KNN mudah dipahami dan diimplementasikan. Model tidak memerlukan pelatihan dalam arti konvensional, karena klasifikasi dilakukan berdasarkan kedekatan jarak dengan data latih.
- Cocok untuk baseline: KNN sering digunakan sebagai baseline untuk membandingkan performa dengan model klasifikasi lain karena sifatnya yang langsung mencerminkan struktur data.

- Tidak mengasumsikan distribusi data tertentu: KNN tidak mengharuskan data memenuhi asumsi distribusi tertentu (seperti normalitas), sehingga fleksibel digunakan pada berbagai jenis data.
- Respon terhadap outlier dapat dikendalikan dengan pemilihan nilai *k* yang tepat.



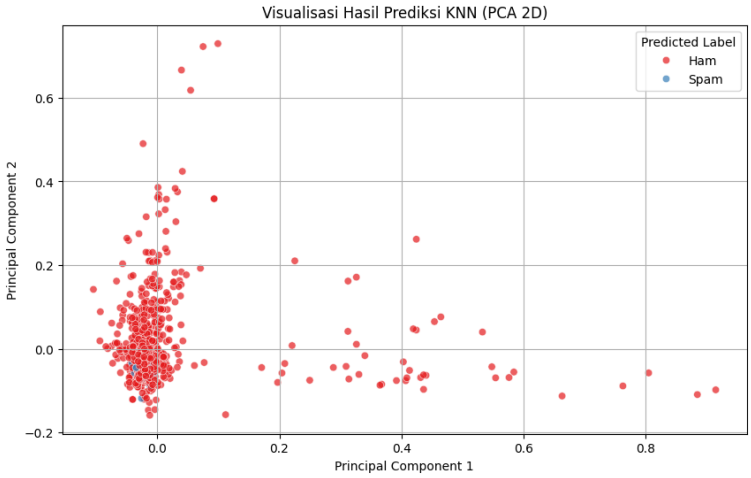
Gambar 12. Modelling KNN

KNN bukanlah algoritma berbasis struktur (seperti pohon keputusan), sehingga tidak menghasilkan diagram alur atau node seperti decision tree. Namun, untuk memahami bagaimana model memisahkan data, dilakukan reduksi dimensi menggunakan PCA (Principal Component Analysis) dari hasil vektorisasi TF-IDF ke dalam 2 dimensi utama.

Visualisasi scatter plot hasil PCA digunakan untuk memetakan prediksi model terhadap dua kelas utama (ham dan spam). Hasil ini memberikan gambaran bagaimana data terkelompok berdasarkan jarak dalam ruang vector:

- Titik-titik berwarna merah atau biru mewakili hasil klasifikasi.
- Penyebaran titik memperlihatkan klasterisasi alami antara kelas spam dan ham berdasarkan fitur yang diekstraksi.

Visualisasi ini membantu memvalidasi secara visual bahwa model mampu membedakan pola pada data dengan cukup baik, meskipun terdapat tumpang tindih pada sebagian area karena keterbatasan representasi 2 dimensi dari data berdimensi tinggi (3000 fitur TF-IDF).

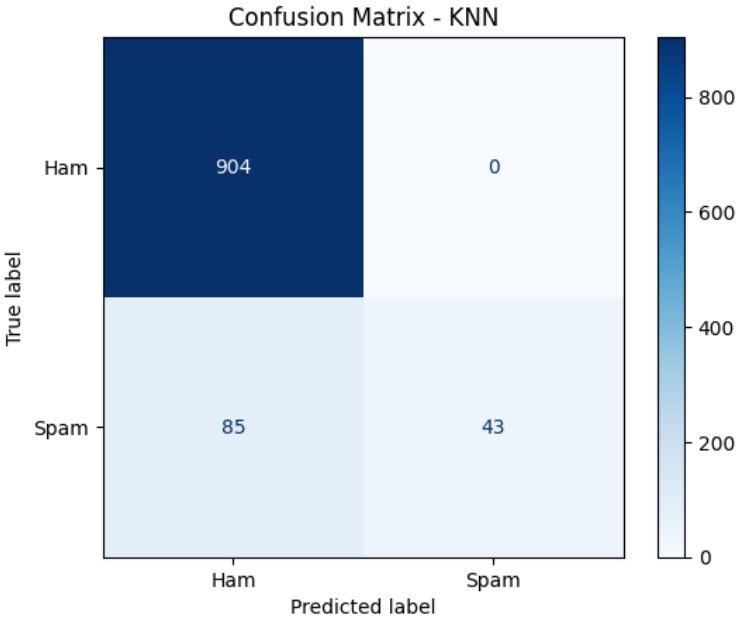


Gambar 13. Visualisasi Hasil prediksi KNN (PCA)

6. Evaluation

6.1 Confusion matrix

Gambar 14. menunjukkan confusion matrix hasil prediksi model KNN terhadap data testing:



Gambar 14. Cunfusion Matrix

- True Negative (TN) = 904: Pesan ham berhasil dikenali dengan benar.
- False Positive (FP) = 0**:** Tidak ada pesan ham yang salah diklasifikasikan sebagai spam.
- False Negative (FN) = 85: Terdapat 85 pesan spam yang tidak terdeteksi (terklasifikasi sebagai ham).
- True Positive (TP) = 43: Pesan spam yang dikenali dengan benar

6.2 Metrik evaluasi

```
=== Classification Report ===
              precision    recall  f1-score   support

         Ham       0.91      1.00      0.96       904
        Spam       1.00      0.34      0.50       128

    accuracy                           0.92      1032
   macro avg       0.96      0.67      0.73      1032
weighted avg       0.92      0.92      0.90      1032
```

Gambar 15. Metrik Evaluasi

Hasil evaluasi menunjukkan bahwa:

- Model memiliki akurasi tinggi (91.8%)**,** namun hal ini kurang representatif karena data tidak seimbang.
- Precision 100% untuk spam menunjukkan bahwa semua pesan yang diprediksi sebagai spam memang benar-benar spam (tidak ada false positive).
- Namun, recall sangat rendah (33.6%)**,** artinya hanya sekitar sepertiga dari pesan spam yang berhasil dideteksi — sisanya lolos sebagai ham.

- F1-score rendah (50.3%)**,** mengindikasikan ketidakseimbangan antara keberhasilan deteksi dan ketepatan deteksi pada kelas spam.

7. Kesimpulan dan Rekomendasi

Proyek ini berhasil membangun model klasifikasi spam SMS menggunakan algoritma K-Nearest Neighbors (KNN) dengan representasi fitur berbasis TF-IDF dan beberapa fitur numerik tambahan. Model menunjukkan akurasi yang tinggi sebesar 91.8% dan precision 100% untuk kelas spam, namun memiliki kelemahan dalam aspek recall yang rendah (33.6%), yang menunjukkan bahwa sebagian besar spam masih gagal terdeteksi. Hal ini menunjukkan bahwa meskipun model mampu meminimalkan false positive, namun masih kurang efektif dalam menangkap spam secara menyeluruh. Oleh karena itu, disarankan agar penelitian selanjutnya menggunakan dataset yang lebih besar dan seimbang, menerapkan teknik balancing data seperti SMOTE, serta membandingkan performa KNN dengan algoritma lain seperti Naive Bayes, SVM, atau Random Forest untuk mendapatkan hasil klasifikasi yang lebih optimal, khususnya pada kelas minoritas seperti spam

8. Referensi

Apriansyah, F. A., Hermawan, A., & Avianto, D. (2024). Optimization of K Value in KNN Algorithm for Spam and HAM Classification in SMS Texts. *International Journal Software Engineering and Computer Science (IJSECS)*, *4*(2), 767–779. https://doi.org/10.35870/ijsecs.v4i2.2681

Arisona, D. C., Wibowo, G. N. A., Siswanto, S., & ... (2023). Klasifikasi Pesan Biasa, Operator, Spam, dan Debt Collector Menggunakan K-Nearest Neighbor. docx. *Jurnal INSYPRO ...*, *8*(November), 1–6. https://journal.uin-alauddin.ac.id/index.php/insypro/article/view/41264%0Ahttps://journal.uin-alauddin.ac.id/index.php/insypro/article/download/41264/18442

Pramakrisna, F. D., Adhinata, F. D., & Tanjung, N. A. F. (2022). Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression. *Teknika*, *11*(2), 90–97. https://doi.org/10.34148/teknika.v11i2.466

Reviantika, F. (2021). Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression. *Jurnal Sistem Cerdas*, *4*(3), 155–160. https://doi.org/10.37396/jsc.v4i3.166

Wara Putera, A., & Dewi Lestari, Y. (2023). Klasifikasi SMS Spam Menggunakan Algoritma K-Nearest Neighbor. *Jikstra*, *5*(01).