JOURNAL SYNTAX IDEA
p–ISSN: 2723-4339 e-ISSN: 2548-1398
Vol. 6, No. 3, Maret 2024

Deteksi Email Spam Dan Non-Spam Berdasarkan Isi Konten Menggunakan Metode *K-Nearest Neighbor* Dan *Support Vector Machine*

Ade Andryani[1], Axel Natanael Salim[2], Tata Sutabri[3]
Magister Teknik Informatika, Universitas Bina Darma
Email: [1]adeanri84@gmail.com, [2]axelsanti610@gmail.com, [3]tata.sutabri@gmail.com

Abstrak

Terhadap banyaknya kasus penyalahgunaan email yang berpotensi merugikan orang lain. Email yang disalahgunakan ini biasa dikenal sebagai email *spam* yang mana email tersebut berisikan iklan, *scam*, bahkan *malware*. Penelitian ini bertujuan untuk mendeteksi email *spam* dan *non-spam* berdasarkan isi konten menggunakan metode *K-Nearest Neighbor* dan *Support Vector Machine* nilai terbaik dari algoritma *K-Nearest Neighbor* dengan pengukuran jarak *Euclidean Distance*. *Support Vector Machine* dan *K-Nearest Neighbor* dapat mengklasifikasi dan mendeteksi *spam* email atau *non-spam* email, *K-Nearest Neighbor* menggunakan perhitungan jarak *Euclidean Distance* dengan nilai K = 1,3, dan 5. Hasil evaluasi menggunakan *confusion matrix* yang menghasilkan bahwa motode *K-Nearest Neighbor* dengan nilai k=3 mendapatkan tingkat akurasi sebesar 92%, tingkat presisi sebesar 91%, *recall* sebesar 100%, dan *f1_score* sebesar 95%. Metode *Support Vector Machine* mendapatkan nilai akurasi sebesar 97% dengan tingkat akurasi sebesar 97%, *recall* sebesar 100%, dan *f1_score* sebesar 98%. Hal ini menjadikan metode *Support Vector Machine* lebih unggal dibandingkan metode *K-Nearest Neighbor* dalam penelitian ini. Selain itu model yang dibangun juga sudah dapat digunakan untuk memprediksi *spam* dan *non spam* dari isi konten email baru.

Kata Kunci: *Confusion Matrix*, Email, KNN, *Spam*, SVM

Abstract

Facing many email problems that have the potential to harm others. This abused email is commonly known as spam email, where the email contains advertisements, scams, and even malware. This study uses the K-Nearest Neighbor method and the Support Vector Machine to detect spam and non-spam emails based on content. The best value of the K-Nearest Neighbor algorithm is determined using Euclidean Distance measurements. Support Vector Machine and K-Nearest Neighbor can classify and detect spam or non-spam emails. K-Nearest Neighbor uses Euclidean Distance calculations with values K = 1, 3, and 5. The evaluation results use a confusion matrix, which shows that the K-Nearest Neighbor method with a value of k=3 achieves an accuracy rate of 92%, a precision level of 91%, a recall of 100%, and an F1-score of 95%. On the other hand, the Support Vector Machine method achieves an accuracy value of 97%, a recall of 100%, and an F1-score of 98%.

How to cite: Ade Andryani[1], Axel Natanael Salim[2], Tata Sutabri[3] (2024) Deteksi Email Spam dan Non-Spam Berdasarkan Isi Konten Menggunakan Metode *K-Nearest Neighbor* dan *Support Vector Machine*, *(6) 3*, https://doi.org/
E-ISSN: 2684-883X
Published by: Ridwan Institute

This makes the Support Vector Machine method superior to the K-Nearest Neighbor method in this study. In addition, the model built can also be used to predict spam and non-spam from the contents of new e-mail content.

Keywords: *Confusion Matrix, Email, KNN, Spam, SVM*

PENDAHULUAN

Perkembangan teknologi telah berkembang sangat pesat yang membawa dunia ke era digital yang serba instan (Megawati, 2021) Perkembangan teknologi internet sekarang berkembang sangat pesat, banyak manfaat yang dirasakan oleh pengguna dengan perkembangan teknologi internet, tetapi banyak pula kejahatan yang dilakukan menggunakan teknologi internet (Laksono et al., 2020). *With the rapid growth of computer and network systems in recent years, there has also been a corresponding increase in cyber-crime* (Wang, 2007) *The number of victims in this type of crime is partly triggered by low public awareness* (Setyawan et al., 2023) Bentuk kejahatan siber berupa penyerangan terhadap pengguna e-mail dengan cara mengirimkan *spam*. *Email is one of the most widely used ways to communicate, with millions of people and businesses relying on it to communicate and share knowledge and information on a daily basis. Nevertheless, the rise in email users has occurred a dramatic increase in spam emails in recent years* (Zavrak & Yilmaz, 2023) *Attackers are becoming more skilled in recent years, using sophisticated technology to produce look-alike emails that make it difficult to distinguish between real and fake ones* (Sibi Chakkaravarthy et al., n.d.) Pesan *spam* pada email ini mengirimkan salinan pesan-pesan yang sama untuk memaksa agar pesan-pesan tersebut sampai kepada pemakai yang tidak mau menerima e-mail tersebut, akibatnya banyak pengguna merasa terganggu oleh banyaknya waktu yang dihabiskan untuk menghapus pesan *spam*, dan besarnya bandwidth jaringan (Sulaeman et al., 2022). *Spam emails are unwanted and unsolicited messages. The major problems in email spam detection methods are low detection rates and a high likelihood of false alarms. This study proposes a hybrid correlation-based deep learning model for email spam classification using a fuzzy inference system* (Ayo et al., 2024)

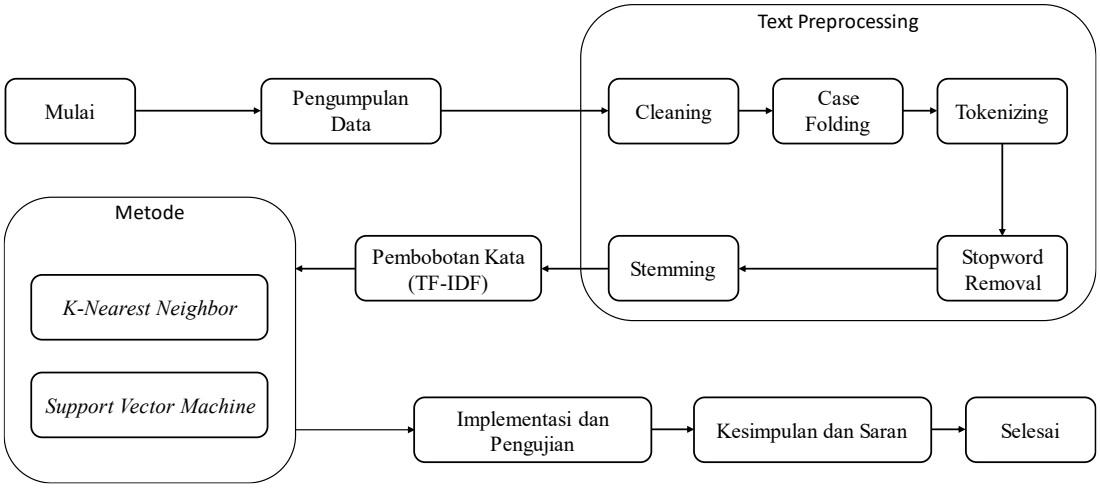
Through digitization as well as globalization, communication in the workplace has changed massively, and email communication is nowadays one important—if not the most important—communication tool. Many people at work, especially managers, feel overwhelmed by the sheer volume and content of the emails that they have to handle (Letmathe & Noll, 2024) Dalam mengatasi masalah tersebut, diperlukan upaya untuk menyaring konten yang masuk ke e-mail pengguna. Penanganan terkait *spam* e-mail telah dilakukan pada penelitian terdahulu. *K-medoids clustering is used for fault detection and classification, while weighted k-Nearest neighbor regression is used to locate the fault* (Gangwar & Shaik, 2023) Penelitian yang dilakukan oleh Laksono, dkk dengan judul "Optimasi Nilai K pada Algoritma KNN untuk Klasifikasi Spam dan Ham Email" bertujuan untuk mengetahui pengklasifikasian *spam* dan *ham* pada email menggunakan metode KNN untuk mengurangi jumlah *spam* dengan cara melakukan

pengecekan menggunakan pendekatan nilai K yang berbeda. Hasil dari penelitian ini diketahui bahwa evaluasi klasifikasi menggunakan *confusion matrix* menggunakan KNN dengan nilai K = 1 memiliki nilai akurasi paling tinggi sebesar 91.4%, dengan menggunakan distribusi frekuensi clustering menghasilkan akurasi yang tinggi sebesar 100%, sedangkan k-means clustering mengahasilkan akurasi sebesar 99% (Laksono et al., 2020). Adapula penelitian yang dilakukan oleh (Hengki & Wahyudi, 2020) dengan judul "Klasifikasi Algoritma Naïve Bayes dan SVM Berbasis PSO Dalam Memprediksi Spam Email Pada Hotline-Sapto". *The Support Vector Machine (SVM) method is one of the popular machine learning algorithms as it gives high accuracy. However, like most machine learning algorithms, the resource consumption of the SVM algorithm in terms of time and memory increases linearly as the dataset grows* (Mutlu & Acı, 2022) Dalam penelitian ini dilakukan pengujian model dengan menggunakan Support Vector Machine (SVM) dan Support Vector Machine (SVM) berbasis Particle Swarm Optimazation (PSO) serta Naïve Bayes dan Naïve bayes berbasis Particle Swarm Optimization (PSO), dari model SVM menghasilkan tingkat accuracy sebesar 84.59% dengan nilai AUC 0.792, untuk pengujian model SVM berbasis PSO menghasilkan tingkat akurasi sebesar 85.25% dengan nilai AUC 0.892, sedangkan pengujian model Naïve Bayes menghasilkan tingkat akurasi sebesar 80.59% dengan nilai AUC 0.942, untuk pengujian model Naïve Bayes berbasis PSO menghasilkan tingkat akurasi sebesar 81.24% dengan nilai AUC sebesar 0.892. Sehingga dapat disimpulkan bahwa model SVM berbasis PSO lebih baik dari pada model lainnya (Hengki & Wahyudi, 2020).

Pengunaan metode yang berbeda membuat hasil yang berbeda pula. Oleh karena itu, penelitian ini melakukan perbandingan metode *Support Vector Machine* (SVM) dengan *K-Nearest Neighbor* (KNN) untuk mencari metode yang optimal dalam membedakan *spam* dan *non spam* pada e-mail. Metode yang digunakan adalah *K-Nearest Neighbor* (KNN) dengan perhitungan jarak *Euclidean* dengan percobaan K = 1,3 dan 5. Kedua metode yang digunakan menggunakan *confusion matrix* untuk evaluasi nilai hasil.

METODE PENELITIAN

Bagian ini akan membahas mengenai langkah-langkah yang digunakan dalam penelitin ini. Langkah pertama yaitu pengumpulan data, lalu *preprocessing*, pembobotan kata, penerapan metode, setelah itu evaluasi, dan terakhir deteksi konten. Gambar 1 merupakan ilustrasi tentang alur yang dilakukan pada penelitian ini.



Gambar 1
***Flowchart* Metodologi Penelitian**

A. Pengumpulan Data

Adapun teknik pengumpulan data merupakan sutau cara yang dilakukan oleh peneliti untuk memperoleh data-data yang diperlukan. Sumber data pada penelitian ini menggunakan data sekunder dimana data penelitian ini menggunakan *dataset spam* dan *non spam* yang berasal dari https://www.kaggle.com/ yang diakses pada tanggal 01 Agustus 2023. *Dataset* yang digunakan terdapat 5572 data *spam* maupun ham(*non spam)*, data dapat dilihat pada tabel 1 sebagai berikut:

Tabel 1 *Dataset Spam Email*

NO	KOMENTAR	CLASS
1	U 447801259231 have a secret admirer who is looking 2 make contact with U-find out who they R*reveal who thinks UR so special-call on 09058094597	Spam
2	Oh k...I'm watching here:)	Ham
3	URGENT! You have won a 1 week FREE membership in our	Spam
4	SMS SERVICES. For your inclusive text credits, please go to www.comuk.net login= 3qxj9 and unsubscribe with STOP at no extra charge. Help 08702840625.COMUK. 220-CM2 9AE	Spam
...	...	...
5570	You are a great role model. You are giving so much, and I wish each day for a miracle, but God has a reason for everything, and I must say I wish I knew why, but I don't. I've looked up to you since I was young, and I still do. Have a great day.	Ham
5571	You are awarded a SiPix Digital Camera! Call 09061221061	Spam

	from landline. Delivery within 28 days. T Cs Box177. M221BP. 2yr warranty. 150ppm. 16. p p3.99	
5572	Yes, princess! I want to please you every night. Your wish is my command...	Ham
5573	Marvel Mobile Play the official Ultimate Spider-man game (4.50) on our mobile right now. Text SPIDER to 83338 for the game & we'll send you a FREE 8 Ball wallpaper.	Spam

B. Text Preprocessing

Text Processing adalah proses dalam membersihkan data sebelum diolah. Pada tahapan ini terdapat 4 proses yaitu :

1. *Case Folding*, pada tahap ini dilakukan penyeragaman teks menjadi huruf kecil, proses *case folding* dapat dilihat pada tabel 2.

Tabel 2 Contoh Proses *Case Folding*

Sebelum *Case Folding*	Free entry in 2 a wkly comp to Win FA final tkts 21st May 2005 Text FA to 87121 to receive entry questions txt rateTCs apply 08452810075over18s
Sesudah *Case Folding*	free entry in 2 a wkly comp to win a final tkts 21st may 2005 text fa to 87121 to receive entry questions txt rates apply 08452810075over18s

2. *Tokenizing*, pada tahap ini dilakukan pemecahan kata pada kalimat, proses *tokenizing* dapat dilihat pada tabel 3.

Tabel 3 Contoh Proses *Tokenizing*

Sebelum *Tokenizing*	free entry in 2 a wkly comp to win a final tkts 21st may 2005 text fa to 87121 to receive entry questions txt rates apply 08452810075over18s
Sesudah *Tokenizing*	[free, entry, in, 2, a, wkly, comp, too, win, fa, cup, final, tkts, 21st, May 2005, text, fa, to, 87121, too, receive, entry, questions, txt, rates, apply, 0845810075over18s]

3. *Stopword Removal*, pada tahap ini dilakukan penghilangan kata yang termasuk kedalam kategori *stopword*, *stopword* merupakan kata yang sering muncul namun dianggap tidak memiliki arti, proses *stopword removal* dapat dilihat pada tabel 4.

Tabel 4 Contoh Proses *Stopword Removal*

Sebelum *Stopword Removal*	[free, entry, in, 2, a, wkly, comp, too, win, fa, cup, final, tkts, 21st, May 2005, text, fa, to, 87121, too, receive, entry, questions, txt,

	rates, apply, 0845810075over18s]
Sesudah *Stopword Removal*	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questions, txt, rates, apply, 0845810075over18s]

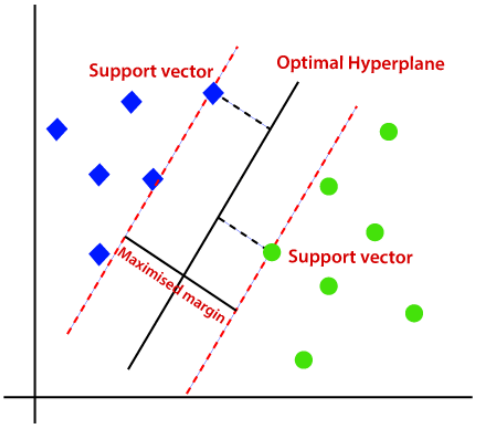
4. *Stemming*, pada tahap ini dilakukan untuk menemukan kata dasar dengan menghilangkan semua kata imbuhan atau kata penghubung, proses *stemming* dapat dilihat pada tabel 5.

Tabel 5 Contoh Proses *Stemming*

Sebelum *Stemming*	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questions, txt, rates, apply, 0845810075over18s]
Sesudah *Stemming*	[free, entry, 2, wkli, comp, win, fa, cup, final, tkt, 21st, may, 2005, text, fa, 87121, receive, entri, questions, txt, ratetcs, apply, 0845810075over18]

a. Support Vector Machine

Support Vector Machine adalah satu dari banyaknya metode regresi atau pengklasifikasi data yang menggunakan data dari data-data sebelumnya dan pemodelannya disuprevisi terlebih dahulu. *Support Vector Machine* menggunakan batas kemampuan yang akan menentukan klasifikasi dari data latih sehingga terbentuk sebuah model *linear* yang paling optimal dalam mengklasifikasi data. Dalam menyelesaikan permasalahan *non-linear* digunakan konsep kernel pada ruang kerja berdimensi tinggi, dengan mencari *hyperplane* yang dapat memaksimalkan margin antar kelas data (Mutawalli et al., 2019). Gambar 2 merupakan gambaran pengklasifikasian menggunakan *Support Vector machine* model *linear*. Penggunaan metode Support Vector Machine betujuan untuk klasifikasi teks dengan menggunakan bobot indeks *term* sebagai fitur (Athira Luqyana et al., 2018).



Gambar 2

Support Vector Machine* Model *Linear

b. K-Nearest Neighbor

K-Nearest Neighbor merupakan salah satu pendekatan yang sederhana untuk diimplmentasikan dan merupakan metode lama yang digunakan dalam pengkasifikasian. *K-Nearest Neighbor* mempunyai tingkat efisiensi yang tinggi dalam beberapa kasus menunjukkan tingkat akurasi yang tinggi dalam hal pengkasifikasian (Asiyah & Fithriasari, 2016). Performa klaksifikasi *K-Nearest Neighbor* dapat dipengaruhi oleh beberapa hal, seperti pemilihan nilai K, pemilihan ukuran jarak, dan sebagainya (Zhang et al., 2017). Dekat atau jauhnya bisa dihutung dengan besaran jarak, dalam penelitian ini menggunakan jarak *Euclidean*. *Euclidean* adalah besarnya jarak suatu garis lurus yang menghubungkan antara objek. Jarak *Euclidean* digunakan karena data teks yang ada sudah mengalami perubahan menjadi angka pada saat *preprocessing* (Laksono et al., 2020). Rumus jarak *Euclidean* dapat dilihat pada persamaan 1 sebagai berikut:

$$d(x_{i,}x_j) = \sqrt{\sum_{n=1}^{P}(x_{ip} - x_{jp})^2} \quad (1)$$

Dengan:

x_{ip} = Data testing ke-i pada variabel ke-p

x_{jp} = Data training ke-j pada variabel ke-p

$d(x_i, x_j)$ = Jarak *euclidean*

P = Dimensi data variabel bebas

c. Evaluasi Model

Evaluasi pada pembelajaran mesin ini digunakan untuk mengukur performa model klasifikasi pada *dataset* yang sudah diketahui labelnya yang dikenal dengan *Confusion Matrix*. Menurut (Aldean et al., 2022) *Confusion Matrix* merupakan cara untuk menunjukkan seberapa baik model yang digunakan dapat mengklasifikasi setiap label yang ada pada *dataset*. Ilustrasi dari *confusion matrix* dapat dilihat pada tabel 6 sebagai berikut.

Tabel 6 Tabel *Confusion Matrix*

	Predicted Class	
	(1) Positive	(0) Negative

Actual Class	(1) Positive	TP (True Positive)	FN (False Negative)
	(0) Negative	FP (False Positive)	TN (True Negative)

Sumber*:* (Afdhal et al., 2022)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (2)$$

$$\text{Precission} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1-Score} = \frac{2\ xRecall\ x\ Precission}{Recall+Precission} \quad (5)$$

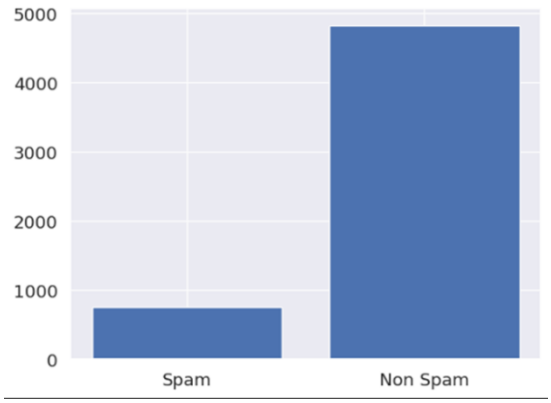
Confusion matrix terdiri dari empat kotak yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Dari empat kotak tersebut, dapat dihitung berapa matrik seperti tingkat akurasi (*accuracy*), presisi (*precision*), *recall* (*sensitivity*), dan *f1-score.* Akurasi adalah presentase jumlah prediksi yang benar dari total data, presisi adalah presentase data *positif* yang benar diklasifikasikan, recall adalah presentase data *positif* yang terdeteksi benar oleh model, dan *f1-score* adalah rata-rata harmonik antara presisi dan *recall* (Santoso et al., 2023).

d. Implementasi dan Pengujian

Implementasi yang dilakukan pada penelitian ini menggunakan Google Colab dengan bahasa pemrograman Python. Tahap pengujian dilakukan untuk menguji kinerja dari mesin yang telah dibangun, untuk menghitung akurasi dan mengidentifikasi algoritma *K-Nearset Neighbor* dan *Support Vector Machine* yang menggunakan *confusion matrix* untuk menghitung tingkat akurasi, tingkat presisi, *recall*, dan *f1-Score*.

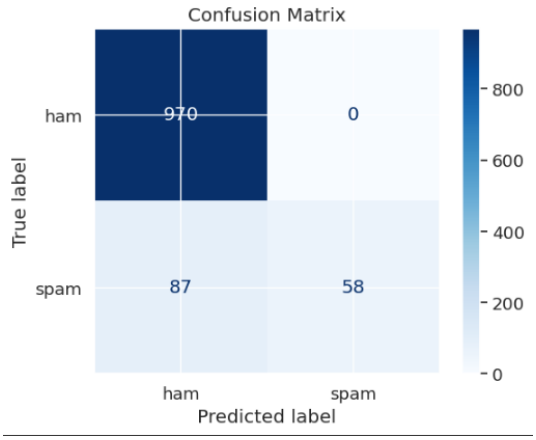
HASIL DAN PEMBAHASAN

Dari data yang digunakan, terdapat 5572 konten spam email yang terdiri dari 4825 email yang termasuk *non spam* dan 747 email yang termasuk email *spam,* sebagaimana yang ditampilkan pada gambar 3.

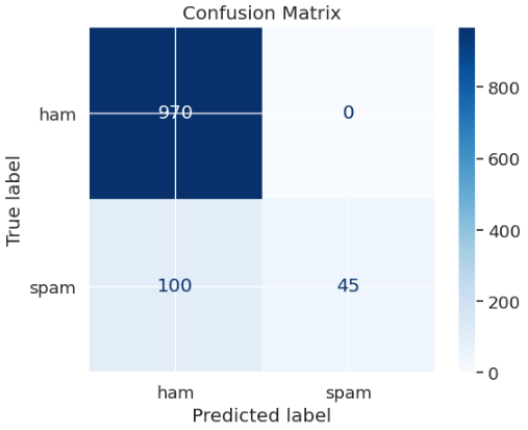


Gambar 3
Perbandingan Class Pada *Dataset*

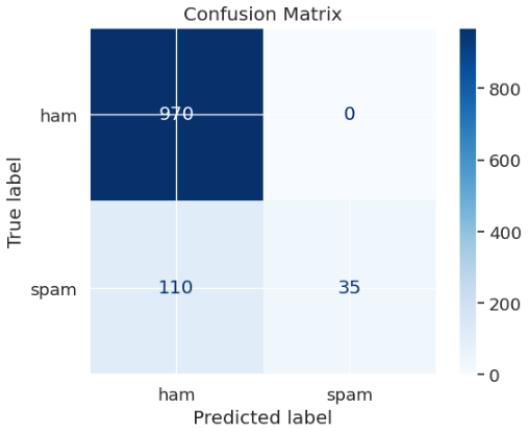
Pada gambar 4 hingga gambar 7 merupakan tampilan visualisasi hasil dari klasifikasi dalam bentuk *heatmap* untuk mempermudah dalam melihat pola TP (*True Positive*), FP (*False Positive*), TN (*True Negative*), TF (*True Negative*) (Sutabri et al., 2018) dari data testing.



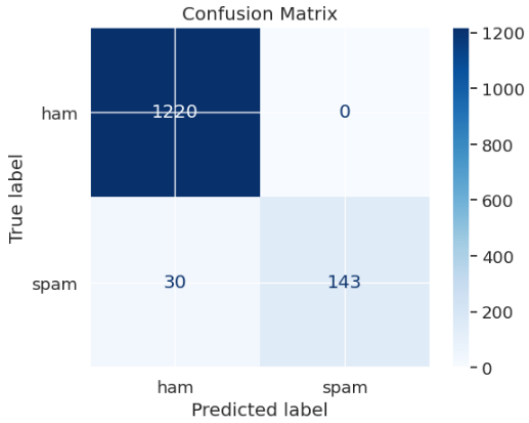
Gambar 4
***Confusion matrix* menggunakan Euclidean dengan k=3**



Gambar 5
***Confusion matrix* menggunakan *Euclidean* dengan k=5**



Gambar 6
***Confusion matrix* menggunakan *Euclidean* dengan k=7**



Gambar 7
***Confusion matrix* dengan metode SVM**

Setelah didapatkan TP (*True Positive*), FP (*False Positive*), TN (*True Negative*), TF (*True Negative*) dari visualisasi diatas, selanjutnya akan dilakukan perhitungan performa dari metode *Support Vector Machine* dan *K-Nearest Neighbor* menggunakan pengukuran jarak *Euclidean* dengan k = 3,5, dan 7. Adapun hasil dari pengujian akan mendapatkan nilai *accuracy*, *precission*, *recall*, dan *f1_score* nya. Tabel 7 menunjukkan hasil pengujian dengan detail nilai *accuracy*, *precision*, *recall*, dan *f1_score* sebagai berikut.

Tabel 7 Tabel Pengujian Data Testing

NO.	Metode			Accuracy	Precision	Recall	F1-Score
1	K-Nearest Neighbor	Euclidean	3	92%	91%	100%	95%
			5	91%	90%	100%	95%
			7	90%	89%	100%	94%
2	Support Vector Machine			97%	97%	100%	98%

Berdasarkan tabel diatas dapat dilihat performa dari metode *K-Nearest Neighbor* dengan pengukuran jarak *Euclidean Distance* dimana k=3 mendapatkan tingkat akurasi sebesar 92%, tingkat presisi sebesar 91%, *recall* sebesar 100%, dan *f1_score* sebesar 95%. Untuk k=5 mendapatkan tingkat akurasi sebesar 91%, tingkat presisi sebesar 90%, *recall* sebesar 100%, dan *f1_score* sebesar 95%. Untuk k=7 mendapatkan tingkat akurasi sebesar 90%, tingkat presisi sebesar 89%, *recall* sebesar 100%, dan *f1_score* sebesar 94%. Metode *Support Vector Machine* mendapatkan nilai akurasi sebesar 97% dengan tingkat akurasi sebesar 97%, *recall* sebesar 100%, dan *f1_score* sebesar 98%. Hal ini menjadikan metode *Support Vector Machine* lebih unggal dibandingkan metode *K-Nearest Neighbor* dalam penelitian ini.

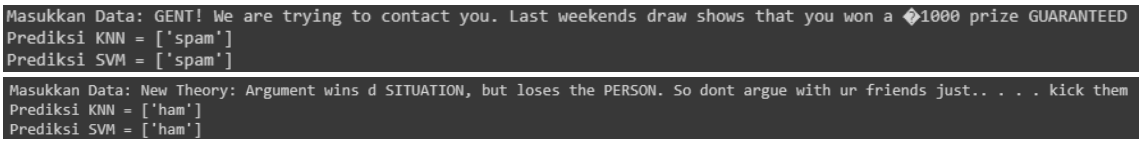
Tingkat akurasi yang tunjukkan oleh metode *Support Vector Machine* ini lebih tinggi dibandingkan dengan *K-Nearest Neighbor* yaitu sebesar 97%, yang artinya metode *Support Vector Machine* ini mampu memprediksi dengan benar terhadap isi konten dengan class pada semua data yang digunakan Selanjutnya metode *Support Vector Machine* juga mendapatkan tingkat *precision* yang lebih tinggi dari pada metode *K-Nearest Neighbor* yaitu sebesar 97%, hal ini menunjukkan bahwa metode *Support Vector Machine* mampu memprediksi label yang sebenarnya dengan benar. Kedua metode yang digunakan menunjukkan persentase *recall* yang sangat baik yaitu 100%, yang artinya kedua metode tersebut sudah sangat baik dalam mengklasifikasi data baik data isi konten email span maupun isi konten email non spam. *F1_score* pada metode *Support Vector Machine* juga memiliki tingkat persentase yang lebih tinggi dari *K-Nearest Neighbor* sehingga metode *Support Vector Machine* lebih baik dalam menggabungkan kemampuan *precision* dan *recall* nya. Selanjutnya akan dilakukan pengujian dari data *testing* yang akan menggunakan kedua metode tersebut, hasil dari pengujian data *testing* dapat dilihat pada tabel 8.

Tabel 8 Hasil Pengujian Data *Testing*

NO.	Email	Class	Pred	
			KNN	SVM
1	U 447801259231 have a secret admirer who is looking 2 make contact with U-find out who they R*reveal who thinks UR so special-call on 09058094597	Spam	Spam	Spam
2	Oh k...i'm watching here:)	Ham	Ham	Ham
3	URGENT! You have won a 1 week FREE membership in our	Spam	Spam	Spam
4	SMS SERVICES. for your inclusive text credits, pls goto www.comuk.net login= 3qxj9 unsubscribe with STOP, no extra charge. help 08702840625.COMUK. 220-CM2 9AE	Spam	Spam	Spam
...	...	...	...	...
5570	You are a great role model. You are giving so much and i really wish each day for a miracle but God as a reason for everything and i must say i wish i knew why but i dont. I've looked up to you since i was young and i still do. Have a great day.	Ham	Ham	Ham
5571	You are awarded a SiPix Digital Camera! call 09061221061 from landline. Delivery within 28days. T Cs Box177. M221BP. 2yr warranty. 150ppm. 16 . p p3.99	Spam	Spam	Spam
5572	Yes princess! I want to please you every night. Your wish is my command...	Ham	Ham	Ham
5573	Marvel Mobile Play the official Ultimate Spider-man game (�4.50) on ur mobile right now. Text SPIDER to	Spam	Spam	Spam

83338 for the game & we ll send u a FREE 8Ball
wallpaper

Selanjutnya dilakukan pengecekan terhadap isi konten email baru yang akan dideteksi, apakah isi konten email tersebut termasuk ke dalam email *spam* atau termasuk ke email *non spam* sebagai tahap akhir penelitian yang dapat dilihat pada gambar 8 sebagai berikut:



Gambar 8
Hasil Deteksi Isi Konten *Spam* dan *Non-Spam*

Dapat dilihat pada gambar diatas, metode yang digunakan baik *K-Nearest Neighbor* maupun *Support Vector Machine* sudah mampu mendeteksi mana isi konten email yang termasuk *spam* maupun isi konten email yang termasuk *non spam* dengan baik.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dan dipaparkan pada hasil dan pembahasan dapat ditarik kesimpulan, yaitu nilai terbaik dari algoritma *K-Nearest Neighbor* dengan pengukuran jarak *Euclidean Distance* dimana k=3 mendapatkan tingkat akurasi sebesar 92%, tingkat presisi sebesar 91%, *recall* sebesar 100%, dan *f1_score* sebesar 95%. Metode *Support Vector Machine* mendapatkan nilai akurasi sebesar 97% dengan tingkat akurasi sebesar 97%, *recall* sebesar 100%, dan *f1_score* sebesar 98%. Hal ini menjadikan metode *Support Vector Machine* lebih unggal dibandingkan metode *K-Nearest Neighbor* dalam penelitian ini. Selain itu model yang dibangun juga sudah dapat digunakan untuk memprediksi *spam* dan *non spam* dari isi konten email baru.

Bibliografi

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *5*(1), 49–54.

Aldean, M. Y., Paradise, P., & Setya Nugraha, N. A. (2022). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac). *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *4*(2), 64–72. https://doi.org/10.20895/inista.v4i2.575

Asiyah, S. N., & Fithriasari, K. (2016). Klasifikasi Berita Online Menggunakan Metode

Support Vector Machine dan K- Nearest Neighbor. *Jurnal Sains Dan Seni ITS*, *5*(2), 317–322.

Athira Luqyana, W., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(11), 4704–4713.

Ayo, F. E., Ogundele, L. A., Olakunle, S., Awotunde, J. B., & Kasali, F. A. (2024). A hybrid correlation-based deep learning model for email spam classification using fuzzy inference system. *Decision Analytics Journal*, *10*, 100390.

Gangwar, A. K., & Shaik, A. G. (2023). k-Nearest neighbour based approach for the protection of distribution network with renewable energy integration. *Electric Power Systems Research*, *220*, 109301.

Hengki, M., & Wahyudi, M. (2020). Klasifikasi Algoritma Naïve Bayes dan SVM Berbasis PSO Dalam Memprediksi Spam Email Pada Hotline-Sapto. *Paradigma - Jurnal Komputer Dan Informatika*, *22*(1), 61–67. https://doi.org/10.31294/p.v22i1.7842

Laksono, E. P., Basuki, A. d, & Abdurrachman Bachtiar, F. (2020). Optimasi Nilai K pada Algoritma KNN untuk klasifikasi Spam dan Ham Email. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(2), 377–383.

Letmathe, P., & Noll, E. (2024). Analysis of email management strategies and their effects on email management performance. *Omega*, *124*, 103002.

Megawati, S. (2021). Pengembangan sistem teknologi internet of things yang perlu dikembangkan negara indonesia. *JIEET (Journal of Information Engineering and Educational Technology)*, *5*(1), 19–26.

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto). *Jurnal Informatika Dan Rekayasa Elektronik*, *2*(2), 43. https://doi.org/10.36595/jire.v2i2.117

Mutlu, G., & Acı, Ç. İ. (2022). SVM-SMO-SGD: A hybrid-parallel support vector machine algorithm using sequential minimal optimization with stochastic gradient descent. *Parallel Computing*, *113*, 102955.

Santoso, H., Putri, R. A., & Sahbandi. (2023). Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest Cyberbullying Comment Detection on Instagram Social Media Using Random Forest Algorithm. *Jurnal Manajemen Informatika (JAMIKA)*, *13*(April), 62–72.

Setyawan, A., Setyabudi, C. M., & Nita, S. (2023). Strategy To Build Public Awareness In Preventing Online Fraud Crimes In The Jurisdiction Of The Cimahi Police. *International Journal of Social Service and Research*, *3*(10), 2641–2649.

Sibi Chakkaravarthy, S., Devi Priya, V. S., Tarun Reddi, M. S. T. R., & Khan, M. K. (n.d.). *A Comprehensive Examination of Email Spoofing: Issues and Prospects for Email Security*.

Sulaeman, Nana Suarna, Abdul Ajiz, Agus Bahtiar, & Fathurrohman. (2022). Perbandingan Kinerja Algoritma Naïve Bayes Dan C.45 Dalam Klasifikasi Spam Email. *KOPERTIP : Jurnal Ilmiah Manajemen Informatika Dan Komputer*, *6*(1), 8–14. https://doi.org/10.32485/kopertip.v6i1.130

Sutabri, T., Suryatno, A., Setiadi, D., & Negara, E. S. (2018). Improving naïve bayes in

sentiment analysis for hotel industry in Indonesia. *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*, 1–6. https://doi.org/10.1109/IAC.2018.8780444

Wang, S.-J. (2007). Measures of retaining digital evidence to prosecute computer-based cyber-crimes. *Computer Standards & Interfaces*, *29*(2), 216–223.

Zavrak, S., & Yilmaz, S. (2023). Email spam detection using hierarchical attention hybrid deep learning method. *Expert Systems with Applications*, *233*, 120977.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3). https://doi.org/10.1145/2990508