Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 13, No. 2, April 2026, hlm. 329-336 e-ISSN: 2528-6579

FITUR INFORMATION GAIN UNTUK MENINGKATKAN NILAI PERFORMA PENGKLASIFIKASI MACHINE LEARNING PADA ANALISIS SENTIMEN KOMENTAR SPAM PENGGUNA YOUTUBE

Gunardi Gunardi[*1], Eni Rohaini[2], Ronald Naibaho[3], Bambang Sukoco[4], Jasmir Jasmir[5]

[1,2,3,4,5]Universitas Dinamika Bangsa, Jambi
Email: [1]gunardi@unama.ac.id, [2]enirohaini0104@gmail.com, [3]rhodes8083@yahoo.co.id,
[4]bengsgkt@gmail.com, [5]ijay_jasmir@yahoo.com
[*]Penulis Korespondensi

(Naskah masuk: 20 Januari 2025, diterima untuk diterbitkan: 15 Desember 2025)

Abstrak

Perkembangan pesat media sosial telah memberikan ruang bagi setiap individu untuk menyampaikan pendapat, baik berupa komentar positif maupun negatif terhadap konten yang mereka akses. Kemudahan dalam memberikan opini secara daring ini berdampak pada semakin besarnya jumlah ulasan yang tersedia. Namun, volume ulasan yang sangat besar sering kali sulit untuk dianalisis secara manual dan berpotensi menimbulkan bias dalam penilaian. Untuk mengatasi permasalahan tersebut, diperlukan pendekatan otomatis melalui klasifikasi sentimen yang bertujuan mengelompokkan opini pengguna ke dalam kategori positif atau negatif. Dalam penelitian ini digunakan tiga algoritma pembelajaran mesin, yaitu Naïve Bayes (NB), K-Nearest Neighbor (KNN), dan R&om Forest (RF). Data penelitian diperoleh dari public dataset UCI Machine Learning. Fokus penelitian adalah meningkatkan kinerja klasifikasi dengan memanfaatkan teknik seleksi fitur information gain. Hasil eksperimen menunjukkan bahwa penerapan information gain secara konsisten meningkatkan performa semua algoritma yang diuji, baik pada metrik akurasi, presisi, recall, maupun f1-score. Naïve Bayes awalnya memperoleh akurasi tertinggi sebesar 74,33% pada kondisi tanpa fitur tambahan. Namun, setelah penerapan information gain, algoritma KNN menunjukkan hasil paling optimal dengan akurasi mencapai 81,28% serta performa yang relatif seimbang pada semua metrik evaluasi. Sementara itu, R&om Forest juga mengalami peningkatan, meskipun tidak melampaui KNN. Secara keseluruhan, penelitian ini menegaskan bahwa pemilihan fitur yang relevan melalui information gain mampu meningkatkan efisiensi dan efektivitas klasifikasi sentimen, serta dapat menjadi pendekatan yang potensial untuk menganalisis opini dalam skala besar.

Kata kunci: *machine learning, information gain, klasifikasi, analisis sentimen, spam*

INFORMATION GAIN FEATURE TO IMPROVE THE PERFORMANCE VALUE OF MACHINE LEARNING CLASSIFICATION ON SENTIMENT ANALYSIS OF YOUTUBE USER SPAM COMMENT

Abstract

The rapid growth of social media has provided individuals with the opportunity to freely express their opinions, whether positive or negative, toward the content they encounter. The increasing ease of sharing opinions online has resulted in a massive volume of user reviews. However, the large number of reviews is difficult to analyze manually & may introduce bias in interpretation. To address this issue, sentiment classification is applied to automatically categorize user opinions into positive or negative classes. In this study, three machine learning algorithms were employed: Naïve Bayes (NB), K-Nearest Neighbor (KNN), & R&om Forest (RF). The dataset was obtained from the public UCI Machine Learning repository. The main objective of this research is to improve classification performance by utilizing feature selection through the information gain method. Experimental results demonstrate that applying information gain consistently enhances the performance of all evaluated algorithms across multiple metrics, including accuracy, precision, recall, & F1-score. Without feature selection, Naïve Bayes achieved the highest accuracy of 74.33%. However, after applying information gain, KNN outperformed the other algorithms by reaching an accuracy of 81.28% & exhibited balanced results across all evaluation metrics. R&om Forest also showed improvement but did not surpass the performance of KNN. Overall, these findings highlight the importance of feature selection in improving both the efficiency & effectiveness of sentiment classification. Furthermore, the use of information gain proves to be a promising approach for large-scale opinion analysis, particularly in h&ling the high dimensionality of textual data.

Keywords: *machine learning, information gain, classification, sentiment analysis, spam*

1. PENDAHULUAN

Pada era informasi dan teknologi saat ini, dan sejak maraknya fasilitas media sosial, opini masyarakat mengalir dengan bebas dan tidak terbatas. Melalui media sosial, setiap orang memiliki hak untuk mengeluarkan pendapat. Hal ini memungkinkan semua hak tersebut diungkapkan melalui *platform* media sosial yang saat ini bisa menyebar dengan cepat dan meluas. *Platform* media sosial paling populer seperti *Instagram, TikTok, Facebook*, dan *Youtube* yang memiliki miliaran pengguna aktif yang sekaligus menjadi pembuat konten (*content creator*) (Farhan Ilham Fadillah, Moch. Alief Chaerobbi, 2025). Selain menjadi *content creator*, sebagian besar *platform* media sosial memungkinkan pengguna memberikan reaksi seperti memberikan komentar (*comment*), juga bereaksi dengan menggunakan fitur menyukai (*like*), tidak menyukai (*dislike*), dan membagikan (*share*)(Sitompul, 2022).

Salah satu *platform* yang dibahas dalam penelitian ini adalah *youtube. Youtube* adalah salah satu *platform* berbagi video bagi pemilik video dan pemirsa dapat melakukannya interaksi seperti *like* atau *dislike* video dan/atau mengomentari video(Harpizon *dkk.*, 2022). Di bagian komentar, pemirsa dapat mengungkapkan pendapat dan emosi terkait dengan video atau bahkan tidak terkait dengan video. Ini juga merupakan cara bagi pemirsa video untuk berinteraksi dengan pembuat video (Umardiyah *dkk.*, 2025). Jumlah yang "*like*" dan persentase "*like*" video, penting bagi pembuat konten di *platform* karena video dengan proporsi ketidaksukaan yang tinggi secara umum memberikan publisitas yang negative (Sitompul, 2022). Dengan sebagian besar data pengguna yang tersedia melalui media sosial, sangat mungkin dapat memperoleh informasi dari pengguna yang mengeluarkan pendapat melalui kata-kata maupun emosi yang meng&ung *spam* (Mathapati *dkk.*, 2016). Dengan banyaknya konten-konten *youtube* dengan berbagai tema, hal ini akan menghasilkan data teks dengan jumlah yang sangat besar yang berasal dari komentar-komentar pengguna yang berinteraksi dengan pembuat konten.

Dengan meningkatnya jumlah data dan kompleksitas data seperti kasus diatas, *machine learning* memberikan penawaran untuk memecahkan masalah ini, dengan kemampuan melakukan proses yang tinggi dan nyaris sempurna. Beberapa metode *machine learning* telah banyak digunakan untuk klasifikasi teks. Sebagai contoh seperti *Naïve Bayes* (Zhang, 2025), *K-Nearest Neighbor* (Jasmir, Nurmaini & Tutuko, 2021), dan *R&om Forest* (Khaleel, Al-Azzawi & Alkhazraji, 2023). *Naive Bayes* sangat sederhana dan efisien serta sangat sensitif terhadap pemilihan fitur(Wasono, 2022). Sementara KNN dikenal dengan kelemahan seperti nilai k yang bias, komputasi yang terlalu komplek, keterbatasan memori, serta mengabaikan atribut yang tidak relevan. Kemudian *R&om Forest* punya kelemahan diantaranya adalah nilai evaluasi bisa berubah secara signifikan dengan hanya perubahan data yang kecil. Pada klasifikasi teks, seleksi fitur berperan penting dalam meningkatkan skalabilitas, efisiensi, dan akurasi proses klasifikasi. Secara umum, metode seleksi fitur yang efektif perlu mempertimbangkan karakteristik domain serta algoritma yang digunakan.

Dengan meluasnya ketersediaan teks dalam format digital dan meningkatnya kebutuhan untuk mengaksesnya secara fleksibel, klasifikasi teks telah menjadi tugas yang fundamental dan penting. Salah satu tantangan utama dalam klasifikasi teks adalah dimensi tinggi dari ruang fitur (Alamin *dkk.*, 2025). yang seringkali terdiri dari puluhan ribu fitur dalam domain teks. Sebagian besar dari fitur ini tidak relevan atau bermanfaat dalam konteks klasifikasi teks, bahkan beberapa di antaranya dapat merugikan akurasi klasifikasi. Selain itu, jumlah besar fitur ini dapat memperlambat proses klasifikasi (Rudolf Huizen, 2023).

Sampai saat ini klasifikasi teks spam masih terus dikembangkan. Penggunaan metode *machine learning* pun menjadi harapan besar dalam penyelesaian masalah teks spam khususnya permodelan dan peningkatan nilai evaluasi kinerja klasifikasi (Azan Rahman, 2021). Namun, karena frekuensi yang tinggi serta jarangnya data teks, penelitian tentang klasifikasi teks memiliki tantangan tersendiri dalam penyelesaiannya (Rahma & Suadaa, 2023). Metode yang optimal dalam menyelesaikan masalah klasifikasi teks adalah metode *machine learning* serta pemilihan fitur yang tepat (Septianingrum & Irawan, 2021)(Kurniabudi, Harris & Ros&a, 2022).

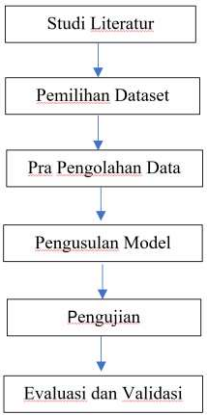
Pemilihan fitur dapat meningkatkan efisiensi dan efektivitas pengklasifikasi(Kurniabudi, Harris & Ros&a, 2022), baik dengan mengurangi jumlah data yang dianalisis maupun mengidentifikasi fitur yang relevan untuk dipertimbangkan dalam proses pembelajaran (Jasmir *dkk.*, 2021). Salah satu fitur yang lebih unggul adalah *information gain* (Kurniabudi *dkk.*, 2020). *Information gain* mengukur sejauh mana kehadiran atau ketiadaan suatu kata memberikan informasi yang berkontribusi pada pengambilan keputusan klasifikasi yang akurat di semua kelas. Information gain merupakan pendekatan filter yang berhasil dalam pengklasifikasi teks.(Perwira *dkk.*, 2022).

Oleh sebab itu, mengacu pada beberapa informasi di atas, maka kami melakukan penelitian untuk meningkatkan akurasi pada beberapa metode *machine learning* yaitu *Naive Bayes, K-Nearest Neighbor* dan *R&om Forest* disertai *Information Gain* sebagai metode pemilihan fitur untuk memperbaiki nilai evaluasi kinerja klasifikasi teks pada komentar spam dari pengguna youtube. Namun *Naive Bayes* masih memiliki kekurangan yaitu saat menghadapi dimensi yang kompleks, maka akan

mengakibatkan tingkat akurasi klasifikasi menjadi rendah dan menghasilkan hasil klasifikasi yang bias(Syahril Dwi Prasetyo, Shofa Shofiah Hilabi & Fitri Nurapriani, 2023). Sementara *K-Nearest Neighbor* memiliki kekurangan diantaranya sangat bergantung pada penskalaan fitur (Utomo, Prabowo & Ju&aputri, 2025). *R&om Forest* memiliki kelemahan, yakni untuk mencapai prediksi dengan tingkat akurasi tinggi, diperlukan sumber daya komputasi yang lebih banyak. Semakin besar kebutuhan akan sumber daya, semakin lama waktu yang diperlukan untuk menghasilkan prediksi (Dara Amelia, 2025).

2. METODE PENELITIAN

Metodologi penelitian ini kami susun dalam bentuk urutan kegiatan atau kerangka kerangka kerja penelitian, seperti yang telihat dalam bagan dibawah ini:



Gambar 1. Kerangka kerja penelitian

Langkah-1: Studi Literatur

Dalam hal ini penulis melakukan studi literatur yang berkaitan dengan analisis sentiment, fitur information gain dan machine learning.

Langkah 2: Pemilihan Dataset

Dataset yang digunakan diambil dari situs UCI Machine Learning dengan link https://archive.ics.uci.edu/ml/datasets/Youtube+Spam+Collection. Dataset ini adalah kumpulan komentar publik yang dikumpulkan untuk penelitian spam. Dataset ini memiliki lima kumpulan data yang disusun oleh 1.956 pesan nyata yang diambil dari lima video yang termasuk di antara 10 video yang paling banyak dilihat pada periode pengumpulan. Informasi datasetnya bisa dilihat pada tabel dibawah ini:

Tabel 1. Informasi Dataset

Dataset	Youtube ID	Spam	Ham	Total
Psy	9bZkp7q19f0	175	175	350
KatyPerry	CevxZvSJLk8	175	175	350
LMFAO	KQ6zr6kCPj8	236	202	438
Eminem	uelHwf8o7_U	245	203	448
Shakira	pRpeEdMmmQ0	174	196	370

Langkah 3. Pra Pengolahan Data

Pada tahap ini data dibersihkan kemudian dilakukan tranformasi data sebelum dilakukan pembuatan model. Tahap preprocessing melalui 3 proses, yaitu: Tokenization, Stopwords Removal dan Stemming.

Langkah 4. Pengusulan Model

Data yang dianalisis kemudian dikelompokkan ke dalam variabel-variabel yang saling berhubungan, dilanjutkan dengan pembuatan model yang sesuai dengan karakteristik data tersebut. Selain itu, pembagian data menjadi data latih (training data) dan data uji (testing data) juga diperlukan dalam proses pengembangan model.

Langkah 5. Pengujian

Bagian ini mengusulkan eksperimen pada model yang akan diuji untuk menghasilkan aturan (rule) yang dapat dimanfaatkan dalam pengambilan keputusan dari hasil penelitian. Proses eksperimen dilakukan menggunakan pemrograman Python, dengan pengujian model dilakukan menggunakan dataset berupa komentar dari YouTube. Pengujian yang kami lakukan terdiri dari 2 eksperimen, eksperimen pertama adalah proses pengujian machine learning tanpa menggunakan fitur, dan eksperimen kedua adalah eksperimen pengujian machine learning dengan menggunakan fitur information gain.

Langkah 6. Evaluasi dan Validasi

Dalam sebuah penelitian, dilakukan evaluasi terhadap model yang digunakan untuk mengukur tingkat akurasi model tersebut dan mendapatkan hasil perb&ingan dari dua model pengujian diatas

3. HASIL DAN PEMBAHASAN

Bagian ini memberikan gambaran umum tentang hasil dan pembahasan yang berasal dari eksperimen yang dilakukan sesuai dengan kerangka penelitian yang diuraikan pada bagian sebelumnya. Eksperimen tersebut berkisar pada penilaian data teks media sosial menggunakan berbagai metode pembelajaran mesin dan fitur information gain. Dengan validasi split 80:20. Pengujian yang dilakukan dalam penelitian ini meliputi pengujian pembelajaran mesin dengan fitur information gain. Pembelajaran Mesin merupakan metode klasifikasi sentimen untuk data teks yang digunakan dalam penelitian ini. Jenis metode pembelajaran mesin yang digunakan adalah: Naive Bayes (NB), K-Nearest Neighbor (KNN) dan R&om Forest (RF).

Tabel 2 merupakan hasil pengujian analisis sentimen komentar spam pengguna youtube yang berjumlah 1956 record menggunakan algoritma Naive Bayes tanpa menggunakan fitur. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Terlihat bahwa hasilnya adalah false positive = 222 dan false negative = 280. Nilai-nilai tersebut tergolong sangat tinggi sehingga menghasilkan nilai akurasi yang rendah.

Tabel 3 merupakan hasil pengujian analisis sentimen komentar spam pengguna youtube yang berjumlah 1956 record menggunakan algoritma Naive Bayes dan menggunakan fitur information

gain. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Terlihat bahwa hasil false positive = 204 dan false negative = 201. Nilai tersebut berada pada area ideal untuk meningkatkan nilai evaluasi kinerja klasifikasi sehingga berdampak pada nilai akurasi yang lebih tinggi.

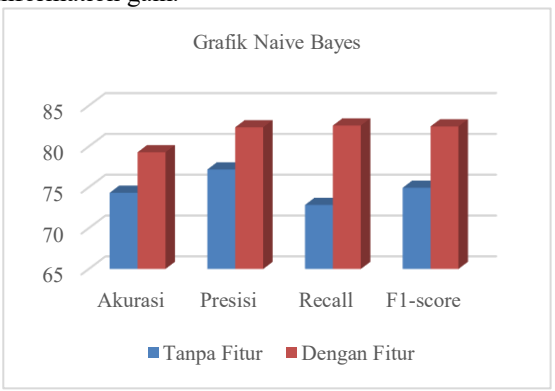
Tabel 2. Confusion Matrix NB tanpa Fitur

		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 751	FP = 222
	Class = No	FN = 280	TN = 703

Tabel 3. Confusion Matrix NB dengan fitur

		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 952	FP = 204
	Class = No	FN = 201	TN = 599

Gambar 2 menggambarkan hasil eksperimen mengenai analisis sentimen komentar spam pengguna youtube yang berjumlah 1956 record. Analisis dilakukan menggunakan teknik Naive Bayes dengan fitur information gain dan tanpa fitur. Pada percobaan ini terlihat adanya peningkatan nilai sebelum menggunakan fitur dan setelah menggunakan fitur information gain.



Gambar 2 Perb&ingan grafik NB tanpa fitur dan dengan fitur

Secara umum dapat dinyatakan bahwa Naive Bayes sering bekerja dengan baik pada data teks karena asumsi independensi kondisionalnya sesuai dengan model representasi kata (Bag-of-Words atau TF-IDF). Namun, saat menggunakan fitur Information Gain, Naive Bayes mungkin tidak sepenuhnya memanfaatkan informasi ini.

Tabel 4 berikut merupakan hasil pengujian analisis sentimen komentar komentar spam pengguna youtube menggunakan algoritma KNN tanpa menggunakan fitur. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Dapat dilihat bahwa hasilnya adalah false positive = 295 dan false negative = 260. Nilai ini tergolong sangat tinggi sehingga menghasilkan nilai akurasi yang rendah.

Tabel 4. Confusion Matrix KNN tanpa fitur

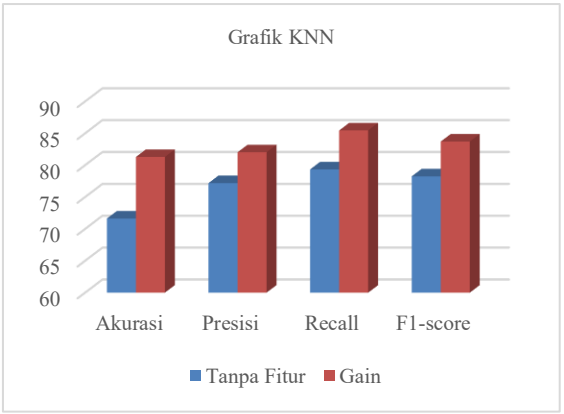
		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 998	FP = 295
	Class = No	FN = 260	TN = 403

Tabel 5 merupakan hasil pengujian analisis sentimen komentar spam pengguna youtube menggunakan algoritma KNN dan menggunakan fitur information gain. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Terlihat bahwa hasil false positive = 206 dan false negative = 160. Nilai tersebut sangat baik untuk meningkatkan nilai evaluasi kinerja klasifikasi, dan cukup signifikan.

Tabel 5. Confusion Matrix KNN dengan Fitur

		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 940	FP = 206
	Class = No	FN = 160	TN = 650

Gambar 3 menjelaskan hasil eksperimen data analisis sentimen komentar spam pengguna youtube sebanyak 900 record, menggunakan metode KNN dengan tiga fitur word embedding dan satu tanpa menggunakan fitur. Pada eksperimen ini terlihat juga terjadi peningkatan nilai sebelum menggunakan fitur dan setelah menggunakan fitur word embedding. Word embedding tertinggi juga dihasilkan oleh fitur FastText. Semua fitur word embedding mampu meningkatkan nilai evaluasi KNN, dan secara umum menghasilkan nilai yang stabil.



Gambar 3. Perb&ingan Grafik KNN tanpa fitur dan dengan fitur

Secara umum dapat dinyatakan bahwa KNN bekerja dengan mencari jarak terpendek antar vektor fitur. Dengan fitur Information Gain, KNN dapat memberikan hasil yang baik jika jarak antar vektor secara efektif memisahkan kelas. Akan tetapi, KNN dapat berjalan lambat dan kurang efisien pada data yang besar karena harus menghitung jarak ke semua titik pada training dataset.

Berikut adalah tabel 6 hasil pengujian analisis sentimen komentar spam pengguna youtube menggunakan algoritma RF tanpa menggunakan fitur. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Dapat dilihat bahwa hasil false positive = 366 dan false negative = 363. Nilai tersebut sangat tinggi sehingga menghasilkan nilai akurasi yang sangat rendah.

Tabel 7 merupakan hasil pengujian analisis sentimen komentar spam pengguna youtube menggunakan algoritma RF dan menggunakan fitur

information gain. Hasil pengujian disimpan dalam matriks konfusi dengan setiap hasil evaluasi. Dapat dilihat bahwa hasilnya adalah false positive = 242 dan false negative = 238. Nilai tersebut sangat baik untuk mendapatkan peningkatan nilai evaluasi kinerja klasifikasi dan menghasilkan nilai terbaik untuk RF.

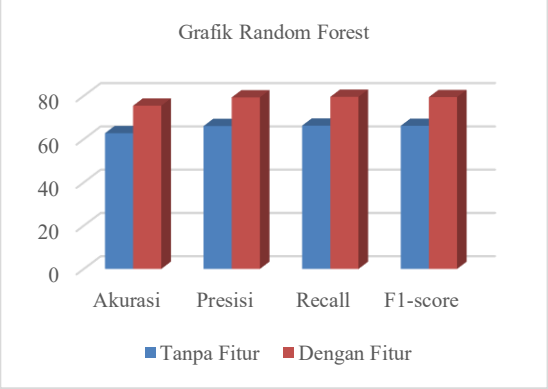
Tabel 6. Confusion Matrix RF tanpa fitur

		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 712	FP = 366
	Class = No	FN = 363	TN = 515

Tabel 7. Confusion Matrix RF dengan Fitur

		Actual Class	
Predicted Class		Class = Yes	Class = No
	Class = Yes	TP = 926	FP = 242
	Class = No	FN = 238	TN = 550

Gambar 4 menjelaskan hasil eksperimen data analisis sentimen komentar spam pengguna youtube sebanyak 1956 record, menggunakan algoritma R&om Forest dengan fitur information gain dan tanpa menggunakan fitur. Pada eksperimen ini terlihat juga terjadi peningkatan nilai sebelum menggunakan fitur dan setelah menggunakan fitur. Fitur information gain mampu meningkatkan nilai evaluasi R&om Forest, dan secara umum menghasilkan nilai yang stabil.



Gambar 4. Perb&ingan Grafik RF tanpa fitur dan dengan fitur

Secara umum dapat dikatakan bahwa R&om Forest cenderung memberikan kinerja yang lebih baik karena memanfaatkan sejumlah besar pohon keputusan dan fitur acak untuk mengurangi overfitting. Dengan fitur information gain, R&om Forest dapat menangkap lebih banyak interaksi antar fitur yang mungkin diabaikan oleh Naive Bayes atau KNN.

Dari seluruh pengujian machine learning yang dilakukan, dapat dilihat bahwa sebelum menggunakan fitur information gain, algoritma Naive Bayes mencapai tingkat akurasi tertinggi sebesar 74,33%, sedangkan algoritma R&om Forest menghasilkan akurasi terendah sebesar 62,73%. Setelah menggunakan fitur information gain, hasil akurasi tertinggi diperoleh dari algoritma KNN dengan nilai sebesar 81,28% dan akurasi terendah diperoleh dari algoritma R&om Forest dengan nilai 75,46%. Setelah melihat seluruh nilai evaluasi kinerja klasifikasi, yaitu akurasi, presisi, recall dan f1-score, algoritma terbaik adalah algoritma KNN dengan hasil evaluasi yang stabil. Seluruh pengujian masih mentoleransi kesalahan false positive dan false negative. Seluruh algoritma masih menggunakan parameter asli. Hal ini dapat menjadi celah untuk penelitian lebih lanjut seperti mengurangi nilai positif palsu atau negatif palsu. Celah untuk meningkatkan akurasi juga dapat dicapai dengan menyetel semua hiperparameter.

Secara umum, hasil ini menegaskan bahwa efektivitas pemilihan fitur berbeda antar algoritma, dengan KNN memperoleh manfaat paling besar. Namun, seluruh algoritma masih menghasilkan tingkat kesalahan (false positive dan false negative) yang cukup tinggi. Oleh karena itu, penelitian lanjutan perlu difokuskan pada optimasi hyperparameter, penerapan teknik praproses lanjutan, dan perluasan dataset agar akurasi serta generalisasi model dapat lebih ditingkatkan.

4. KESIMPULAN

Dalam makalah ini, kami menyelidiki dampak berbagai metode pembelajaran mesin yang dikombinasikan dengan fitur information gain pada kinerja klasifikasi analisis sentimen komentar spam pengguna youtube. Secara khusus, kami memb&ingkan kinerja algoritma Naive Bayes (NB), K-Nearest Neighbor (KNN), dan R&om Forest (RF) sebelum dan sesudah menggabungkan fitur information gain. Eksperimen menunjukkan bahwa semua metode pembelajaran mesin mengalami peningkatan dalam metrik kinerja klasifikasi—akurasi, presisi, ingatan, dan skor F1—ketika fitur information gain diterapkan. Di antara metode yang diuji, NB mencapai akurasi tertinggi sebesar 74,33% tanpa menggunakan fitur. Setelah menggunakan fitur, KNN menghasilkan akurasi tertinggi sebesar 81,28%. Selain itu, KNN menunjukkan kinerja yang seimbang di semua metrik evaluasi (akurasi, presisi, ingatan, dan skor F1) ketika dikombinasikan dengan fitur information gain, yang menggarisbawahi kekokohan dan efektivitasnya dalam tugas analisis sentimen. Temuan ini menyoroti peran penting fitur information gain dalam meningkatkan kinerja algoritme pembelajaran mesin untuk klasifikasi sentimen. Pekerjaan di masa mendatang dapat difokuskan pada pengurangan lebih lanjut rasio positif palsu dan negatif palsu dengan menyempurnakan hiperparameter dan menggunakan teknik praproses yang lebih canggih. Memperluas kumpulan data dengan sampel yang lebih beragam juga dapat meningkatkan ketahanan dan generalisasi model.

DAFTAR PUSTAKA

ALAMIN, Z. *dkk.* 2025. Optimasi Ekstraksi Fitur Citra Karakter Font Menggunakan

Algoritma Support Vector Machines (SVM) untuk Klasifikasi Tipografi. *Scientific : Journal of Computer Science & Informatics*, 2(1), pp. 30–39. Available at: https://doi.org/10.34304/scientific.v2i1.344.

AZAN RAHMAN, A.M. . 2021. ANALISIS KLASIFIKASI EMAIL SPAM MENGGUNAKAN ALGORITMA NAÏVE BAYES. *Jurnal Comasie*, 6(2), pp. 40–51. Available at: http://ejournal.upbatam.ac.id/index.php/comasiejournal%0AJurnal Comasie ISSN (Online) 2715-6265%0APERANCANGAN.

DARA AMELIA, R.K.R. 2025. Penerapan Algoritma R&Om Forest Untuk Prediksi Pejualan Dan Persedian Produk Pada Toko Frozen Food Anisa. *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, 15(2), pp. 20–29.

FARHAN ILHAM FADILLAH, MOCH. ALIEF CHAEROBBI, M.D.P. 2025. Inovasi Konten Video Berita Di Harian Bhirawa Dalam Transformasi Media Cetak Ke Platform Media Youtube', *Prosiding Seminar Nasional Mahasiswa Komunikasi (SEMAKOM)*, 03(02), pp. 77–80.

HARPIZON, H.A.R. *dkk.* 2022. Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes', *... Di YouTube ...*, 5(1), pp. 131–140.

JASMIR, J. *dkk.* 2021. Bigram feature extraction & conditional r&om fields model to improve text classification clinical trial document. *Telkomnika (Telecommunication Computing Electronics & Control)*, 19(3), pp. 886–892. Available at: https://doi.org/https://doi.org/10.12928/TELKOMNIKA.v19i3.18357.

JASMIR, J., NURMAINI, S. & TUTUKO, B. 2021. Fine-grained algorithm for improving knn computational performance on clinical trials text classification. *Big Data & Cognitive Computing*, 5(4). Available at: https://doi.org/https://doi.org/10.3390/bdcc5040060.

KHALEEL, A.A., AL-AZZAWI, A.A.M. & ALKHAZRAJI, A.M. 2023. R&om forest for lung cancer analysis using Apache Mahout & Hadoop based on software defined networking. *Indonesian Journal of Electrical Engineering & Computer Science*, 32(2), pp. 1086–1093. Available at: https://doi.org/https://doi.org/10.11591/ijeecs.v32.i2.pp1086-1093.

KURNIABUDI *dkk.* 2020. CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection. *IEEE Access*, 8, pp. 132911–132921. Available at: https://doi.org/10.1109/ACCESS.2020.3009843.

KURNIABUDI, K., HARRIS, A. & ROS&A, E. 2022. Optimalisasi Seleksi Fitur Untuk Deteksi Serangan Pada IoT Menggunakan Classifier Subset Evaluator', *JURIKOM (Jurnal Riset Komputer)*, 9(4), p. 885. Available at: https://doi.org/10.30865/jurikom.v9i4.4618.

Mathapati, S. *dkk.* 2016. Sentiment Analysis & Opinion Mining from Social Media : A Review A Model of Product Performance Forecasting: A Hybrid. *Global Journal of Computer Science & Technology*, 16(5), pp. 0975–4172.

PERWIRA, R.I. *dkk.* 2022. Effect of information gain on document classification using k-nearest neighbor. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 8(1), pp. 50–57. Available at: https://doi.org/10.26594/REGISTER.V8I1.2397.

RAHMA, I.A. & SUADAA, L.H. 2023. Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(6), pp. 1329–1340. Available at: https://doi.org/10.25126/jtiik.2023107325.

RUDOLF HUIZEN, V.K.N. 2023. Optimalisasi Ekstraksi Fitur dan Klasifikasi untuk Deteksi Objek di IoT. *Jurnal Sistem dan Informatika (JSI)*, 18(1), pp. 74–79. Available at: https://doi.org/10.30864/jsi.v18i1.602.

SEPTIANINGRUM, F. & IRAWAN, A.S.Y. 2021. Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review. *Jurnal Media Informatika Budidarma*, 5(3), p. 799. Available at: https://doi.org/10.30865/mib.v5i3.2983.

SITOMPUL, W.W. 2022. Penelitian Tentang Youtube. *Jurnal Perpustakaan dan Informasi*, 2275.

SYAHRIL DWI PRASETYO, SHOFA SHOFIAH HILABI & FITRI NURAPRIANI 2023. Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, 10, pp. 1–7. Available at: https://doi.org/10.35134/komtekinfo.v10i1.330.

UMARDIYAH, F. *DKK.* 2025. Aktualisasi Kinerja melalui Pelatihan Pembuatan Video Iklan Berbasis Artificial Intellegence untuk Mendukung Program Kerja Karang Taruna. *Jumat Informatika: Jurnal Pengabdian Masyarakat*, 6(1), pp. 11–15. Available at: https://doi.org/10.32764/abdimasif.v6i1.5339.

UTOMO, W.B., PRABOWO, H. & JU&APUTRI,

D.S. 2025. MODEL DETEKSI DINI DIABETES BERBASIS K-Nearest Neighbor', *Journal Of Computer Science & Artificial Intelligence*, 2(1), pp. 1–10.

WASONO, R. 2022. Perb&Ingan Metode R&Om Forest Dan Naïve Bayes Untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit. *Seminar Nasional Edusaintek* [Preprint].

ZHANG, L. 2025. Features extraction based on Naive Bayes algorithm & TF-IDF for news classification. *Plos One*, 20(7 July), pp. 1–17. Available at: https://doi.org/10.1371/journal.pone.0327347.