SKRIPSI

ANALISIS KINERJA ALGORITMA *DECISION TREE* DAN *RANDOM FOREST* PADA KLASIFIKASI MULTIKELAS CUITAN X MENGGUNAKAN *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF)

(Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)

JAMILA MAULIDA SHOLICHATI
NIM. 22106010009

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
YOGYAKARTA
2026

ANALISIS KINERJA ALGORITMA *DECISION TREE* DAN *RANDOM FOREST* PADA KLASIFIKASI MULTIKELAS CUITAN X MENGGUNAKAN *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF)

(Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)

Skripsi

Untuk memenuhi sebagian persyaratan

mencapai derajat Sarjana S-1

Program Studi Matematika

diajukan oleh

JAMILA MAULIDA SHOLICHATI

NIM. 22106010009

Kepada

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN TUGAS AKHIR

Universitas Islam Negeri Sunan Kalijaga FM-UINSK-BM-05-03/R0

SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi / Tugas Akhir
Lamp :

Kepada
Yth. Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta
di Yogyakarta

Assalamu'alaikum wr. wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka kami selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama	: Jamila Maulida Sholichati
NIM	: 22106010009
Judul Skripsi	: Analisis Kinerja Algoritma *Decision Tree* dan *Random Forest* pada Klasifikasi Multikelas Cuitan X Menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) (Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)

sudah dapat diajukan kembali kepada Program Studi Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Matematika.

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara tersebut di atas dapat segera dimunaqasyahkan. Atas perhatiannya kami ucapkan terima kasih.

Wassalamu'alaikum wr. wb.

Yogyakarta, 4 Juni 2026

Pembimbing I	Pembimbing II
Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.	Muhamad Rashif Hilmi, S.Si., M.Sc.
NIP. 19660731 200003 2 001	NIP. 19920309 202012 1 001

HALAMAN PENGESAHAN

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1215/Un.02/DST/PP.00.9/06/2026

Tugas Akhir dengan judul : Analisis Kinerja Algoritma Decision Tree Dan Random Forest Pada Klasifikasi Multikelas Cuitan X Menggunakan Term Frequency Inverse Document Frequency (TF-IDF) (Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)

yang dipersiapkan dan disusun oleh:

Nama : JAMILA MAULIDA SHOLICHATI
Nomor Induk Mahasiswa : 22106010009
Telah diujikan pada : Kamis, 04 Juni 2026
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6a2279c3aa4ab

Penguji I

Dr. Epha Diana Supandi, S.Si., M.Sc.
SIGNED

Valid ID: 6a22764ade628

Penguji II

Muhamad Rashif Hilmi, S.Si., M.Sc.
SIGNED

Valid ID: 6a224facca54a

Yogyakarta, 04 Juni 2026
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6a2279c3a4e91

HALAMAN PERNYATAAN KEASLIAN

SURAT PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	: Jamila Maulida Sholichati
NIM	: 22106010009
Program Studi	: Matematika
Fakultas	: Sains dan Teknologi

Dengan ini menyatakan bahwa isi skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu Perguruan Tinggi dan sesungguhnya skripsi ini merupakan hasil pekerjaan penulis sendiri sepanjang pengetahuan penulis, bukan duplikasi atau saduran dari karya orang lain kecuali bagian tertentu yang penulis ambil sebagai bahan acuan. Apabila terbukti pernyataan ini tidak benar, sepenuhnya menjadi tanggung jawab penulis.

Yogyakarta, 22 Mei 2026

Jamila Maulida Sholichati

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Allah SWT,
karya ini penulis persembahkan untuk:

Kedua orang tua dan keluarga tercinta, atas doa, kasih sayang, dukungan, serta segala pengorbanan yang senantiasa mengiringi langkah penulis.

Diri penulis sendiri, yang telah mampu bertahan, terus berusaha, dan tidak menyerah hingga sampai pada titik ini.

Serta seluruh hal baik yang mengantarkan penulis melalui setiap proses hingga skripsi ini dapat diselesaikan.

HALAMAN MOTTO

فَإِنَّ مَعَ الْعُسْرِ يُسْرًا ۝٥ إِنَّ مَعَ الْعُسْرِ يُسْرًا ۝٦

"Maka, sesungguhnya beserta kesulitan ada kemudahan. Sesungguhnya beserta kesulitan ada kemudahan." (QS. Al-Insyirah : 5-6)

"Bersemangatlah dalam hal yang bermanfaat untukmu, minta tolonglah pada Allah, dan jangan patah semangat." (HR. Muslim)

"*Even if you're not confident, you're still a precious person.*" (Lee Haechan)

"*Don't change, let go. Be brave, be strong. But take it easy, no rush. I know you got it. Go, fail, forward. Go get your voice hard, don't fold.*"
(UNKNOWN - NCT DREAM)

"*I live my life. Now I'm not afraid. Don't hesitate. Life is like a miracle*"
(Live My Life - aespa)

"*Done is better than perfect*"

PRAKATA

Bismillahirrahmanirrahim

Alhamdulillahirabbil'alamin, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul "Analisis Kinerja Algoritma *Decision Tree* dan *Random Forest* pada Klasifikasi Multikelas Cuitan X Menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) (Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)". Shalawat serta salam senantiasa tercurah kepada Nabi Muhammad SAW, suri teladan yang membawa umat manusia menuju jalan penuh ilmu dan keberkahan. Adapun penyusunan skripsi ini sebagai salah satu syarat dalam menyelesaikan studi Sarjana Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta.

Penulis menyadari bahwa proses penyusunan skripsi ini tidak akan terselesaikan tanpa dukungan, bimbingan, kontribusi serta motivasi dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, dengan penuh rasa hormat dan kerendahan hati, penulis menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Noorhaidi Hasan, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor UIN Sunan Kalijaga Yogyakarta.
2. Ibu Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
3. Ibu Dr. Epha Diana Supandi, S.Si., M.Sc., selaku Ketua Program Studi Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta.
4. Ibu Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dosen Pembimbing Akademik dan Dosen Pembimbing Skripsi yang telah memberikan bimbingan, ketelitian, serta arahan kepada penulis selama proses penyusunan skripsi.

5. Bapak Muhamad Rashif Hilmi, S.Si., M.Sc., selaku Dosen Pembimbing Skripsi yang telah meluangkan waktu, memberikan bimbingan dan berbagai masukan yang membangun sehingga penulis dapat memahami dan menyelesaikan penelitian skripsi dengan lebih baik.
6. Seluruh Dosen Prodi Matematika dan seluruh Dosen beserta Staf Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta yang telah memberikan ilmu, pengalaman, serta bantuan selama masa studi penulis.
7. Kedua orang tua penulis, Abi Rispriyanto dan Umi Dwi Astuti yang senantiasa memberikan cinta dan kasih sayang, yang tidak pernah berhenti mendoakan, mendukung, serta memperjuangkan penulis dalam setiap kondisi. Terima kasih telah menjadi tempat pulang dan sumber kekuatan utama bagi penulis untuk menyelesaikan setiap proses hingga titik ini.
8. Adik penulis, Yahya Muhammad yang kehadiran dan kebersamaannya selalu menguatkan penulis dalam menjalani setiap proses. Serta yang tercinta, Almh. Fadhila Rachmawati, adik penulis yang kehadirannya selalu dirindukan dan dalam diam selalu menjadi sumber kekuatan serta alasan penulis untuk menyelesaikan setiap proses, *you'll always be in my heart*.
9. Seluruh kerabat terdekat, saudara, dan keluarga besar yang tidak dapat penulis sebutkan satu per satu, terima kasih selalu turut memberi doa, dukungan, serta kasih sayang yang terus mengalir kepada penulis.
10. Lee Haechan beserta seluruh anggota grup musik NCT, aespa, RIIZE, dan Hearts2Hearts, yang secara tidak langsung menjadi sumber semangat penulis melalui karya-karya dan dedikasi mereka, sehingga mampu memberikan motivasi dan hiburan bagi penulis dalam melewati berbagai proses hingga terselesaikannya skripsi ini.
11. Teman-teman *Komang* (Viga, Tintan, Atina, Bunga, Dzakiyya, Amanda), teman-teman dekat penulis dari awal masa perkuliahan hingga saat ini, terima kasih telah memberi kebersamaan, tawa, dukungan, serta menjadi ruang untuk berbagi sehingga proses perkuliahan penulis terasa lebih ringan.

12. Alya, teman dekat yang telah mengenal penulis selama kurang lebih 10 tahun, terima kasih atas kebersamaan yang terus terjaga, atas dukungan, tawa, dan semua cerita yang dilewati bersama hingga saat ini.
13. Anggi, teman yang banyak membersamai penulis di masa perkuliahan, terima kasih telah menjadi teman yang sefrekuensi dalam banyak hal, memberi dukungan dan momen sederhana yang membantu meringankan penulis dalam menjalani proses ini. Begitu juga Riska, terima kasih telah membersamai dan menjadi teman penulis di masa perkuliahan, turut berbagi cerita dan menjadi bagian dari proses yang memberikan warna bagi penulis.
14. Teman-teman penulis sejak SMP, *Banteng Prindapan* (Fahma, Vika, Eva, Nisfir, Wihda, Jihan, Tita, Zahra, Yaya) yang telah menjadi bagian dari perjalanan panjang penulis hingga saat ini, terima kasih atas hubungan yang tetap terjaga meskipun waktu dan kesibukan terus berjalan, serta atas kehadiran yang selalu memberi arti tersendiri bagi penulis.
15. Teman-teman Matematika angkatan 2022 yang telah membersamai dan memberi dukungan baik dalam perkuliahan, diskusi tugas, maupun kegiatan lainnya, sehingga banyak memberikan pembelajaran bagi penulis selama menempuh masa studi.
16. Teman-teman KKN 117 Purbosono Wonosobo, terima kasih atas kebersamaan, kerja sama, serta pengalaman berharga yang menjadikan masa pengabdian penulis selama kurang lebih 45 hari terasa begitu berkesan dan penuh makna.
17. Teman-teman HM-PS Matematika periode 2023/2024, terutama Departemen Minat dan Bakat yang telah bekerja sama dengan baik, memberikan kesempatan dan pengalaman bagi penulis selama masa kepengurusan.
18. Seluruh pihak yang telah terlibat dan tidak dapat penulis sebutkan satu per satu, terima kasih atas segala bentuk bantuan, dukungan, dan kontribusi yang telah diberikan kepada penulis dalam proses penyusunan skripsi ini.

19. *Last but not least*, penulis ingin berterima kasih kepada diri sendiri karena telah mampu bertahan, terus berusaha, dan tidak menyerah dalam melewati setiap proses hingga akhirnya skripsi ini dapat diselesaikan.

Penulis berharap semoga skripsi ini dapat memberikan manfaat bagi semua pihak yang membacanya. Penulis juga menyadari bahwa skripsi ini masih memiliki keterbatasan, sehingga kritik dan saran yang membangun sangat diharapkan.

Yogyakarta, 18 Mei 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR SIMBOL

a_{ij}	:	elemen matriks pada baris ke-i dan kolom ke-j
$\vec{v}$	:	vektor
$\|\mathbf{v}\|$	:	panjang atau norm vektor
$TF_{t,d}$	:	nilai TF *term* ke-t pada dokumen ke-d
IDF_t	:	nilai IDF *term* ke-t
DF_t	:	jumlah dokumen yang mengandung *term* ke-t
$TFIDF_{t,d}$	:	bobot TF-IDF *term* ke-t pada dokumen ke-d
N	:	jumlah total dokumen atau total data
p_i	:	proporsi data pada kelas ke-i
$Entropy(S)$	:	nilai *entropy* himpunan data S
$IG(S, A)$	:	*information gain* atribut A
S	:	himpunan data
S_v	:	subset data dengan nilai atribut v
$Gini(S)$	:	nilai *gini index*
$R_\alpha(T)$	:	*cost complexity pruning*
D	:	dataset pelatihan
D_b	:	sampel *bootstrap* ke-b
x_i	:	data atau fitur pengamatan ke-i
y_i	:	label kelas pengamatan ke-i
$\hat{y}$	:	hasil prediksi kelas

INTISARI

ANALISIS KINERJA ALGORITMA *DECISION TREE* DAN *RANDOM FOREST* PADA KLASIFIKASI MULTIKELAS CUITAN X MENGGUNAKAN *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF)

(Studi Kasus: Data Cuitan Terkait Grup K-pop aespa)

Oleh

JAMILA MAULIDA SHOLICHATI

22106010009

Klasifikasi teks merupakan proses pengelompokan data teks ke dalam kategori tertentu berdasarkan karakteristik kata atau pola bahasa yang terkandung di dalamnya. Aktivitas pengguna pada media sosial X menghasilkan data teks tidak terstruktur dalam jumlah besar sehingga memerlukan metode klasifikasi untuk mengidentifikasi jenis konten pada cuitan. Cuitan terkait grup K-pop aespa diklasifikasikan ke dalam empat kategori, yaitu Informasi, Opini/Ekspresi, Interaksi Fandom, dan Promosi menggunakan algoritma *Decision Tree* dan *Random Forest* dengan representasi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF). Dataset penelitian terdiri atas 2304 cuitan hasil *scraping* dan pelabelan manual. Tahapan *preprocessing* meliputi *cleaning*, *tokenization*, *stopword removal*, dan *stemming*, kemudian dilakukan ekstraksi fitur menggunakan TF-IDF. Hasil evaluasi menunjukkan bahwa algoritma *Decision Tree* memperoleh akurasi sebesar 70%, sedangkan *Random Forest* memperoleh akurasi sebesar 75%. Hasil tersebut menunjukkan bahwa *Random Forest* memiliki performa lebih baik dibandingkan *Decision Tree* dalam klasifikasi multikelas data cuitan terkait grup aespa pada platform media sosial X.

Kata kunci: klasifikasi teks, *Decision Tree*, *Random Forest*, TF-IDF, media sosial X, multikelas

ABSTRACT

PERFORMANCE ANALYSIS OF DECISION TREE AND RANDOM FOREST FOR MULTICLASS CLASSIFICATION OF X TWEETS USING TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

(Case Study: Tweet Data Related to K-pop Group aespa)

By

JAMILA MAULIDA SHOLICHATI

22106010009

Text classification is the process of grouping text data into specific categories based on the characteristics of the words or language patterns they contain. User activity on social media platform X generates a large amount of unstructured text data, necessitating a classification method to identify the content types of tweets. Tweets related to the K-pop group aespa were classified into four categories: Information, Opinion/Expression, Fandom Interaction, and Promotion using the Decision Tree and Random Forest algorithms with Term Frequency-Inverse Document Frequency (TF-IDF) feature representation. The research dataset consisted of 2,304 tweets scraped and manually labeled. Preprocessing steps included cleaning, tokenization, stopword removal, and stemming, followed by feature extraction using TF-IDF. The evaluation results showed that the Decision Tree algorithm achieved an accuracy of 70%, while the Random Forest algorithm achieved an accuracy of 75%. These results indicate that Random Forest outperformed Decision Tree in the multiclass classification of tweet data related to the group aespa on social media platform X.

Keywords: text classification, Decision Tree, Random Forest, TF-IDF, social media X, multiclass

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang pesat telah menghasilkan jumlah data digital sangat besar dan beragam, selanjutnya dikenal dengan istilah *big data*. *Big data* memiliki karakteristik *volume*, *velocity*, dan *variety*, sehingga metode konvensional dalam pengolahan data seringkali tidak memadai untuk mengekstraksi informasi yang bermakna dari data tersebut (Kitchin, 2014; Zikopoulos et al., 2012). Salah satu bentuk data yang dominan dalam ekosistem *big data* adalah data tidak terstruktur, khususnya data teks. Data ini bersifat tidak baku, informal, dan dapat mengandung kode campuran bahasa, sehingga tidak dapat diproses secara langsung menggunakan metode statistik konvensional (De Boe, 2014; Sint et al., 2009). Sumber data teks berasal dari media sosial karena memungkinkan pengguna menghasilkan konten secara *real-time* dan dalam jumlah besar (Nasrullah, 2021). Karakteristik data teks media sosial yang ringkas, dinamis, dan bervariasi menjadikan *text mining* sebagai pendekatan untuk mengekstraksi pola dan informasi bermakna (Aggarwal & Zhai, 2012; Feldman & Sanger, 2007).

Platform media sosial X dipilih sebagai sumber data utama karena kemampuannya menyediakan konten secara *real-time* dan publik, sehingga memudahkan pengumpulan data dalam jumlah besar tanpa menghadapi banyak pembatasan privasi. Dibandingkan dengan platform media sosial lain seperti Instagram, Facebook, atau TikTok, X memiliki format teks yang ringkas dan konsisten, memfasilitasi proses *tokenization* dan representasi numerik untuk klasifikasi multikelas. Meskipun akses *API* resmi memiliki keterbatasan, alternatif *social media scraping* menggunakan *Tweet Harvest* dapat membantu pengumpulan data secara efisien, termasuk variasi bahasa dan kode campuran yang relevan untuk studi klasifikasi teks. Karakteristik platform X yang kaya interaksi sosial, didukung

oleh *hashtag* maupun *mention*, membuatnya ideal untuk menganalisis topik yang bervariasi dan dinamis, sehingga metode klasifikasi multikelas dapat diterapkan secara efektif untuk mengkategorikan cuitan ke dalam berbagai topik, bukan sekadar analisis sentimen.

Fokus penelitian ini diarahkan pada klasifikasi cuitan di platform media sosial X yang berkaitan dengan salah satu grup musik K-pop yaitu aespa. Pemilihan K-pop sebagai topik penelitian dikarenakan peneliti mengikuti genre musik ini sehingga memiliki pemahaman kontekstual yang lebih baik terhadap konten dan dinamika komunitas. Hal ini memungkinkan proses penentuan kategori untuk klasifikasi multikelas dilakukan secara lebih terarah dan efisien tanpa harus melakukan penelusuran awal. Pemilihan topik K-pop juga didasarkan pada tingginya volume interaksi pengguna dan keragaman konten yang tersedia di platform media sosial, sehingga menyediakan dataset yang kaya dan representatif untuk analisis teks. Pemilihan grup K-pop aespa secara khusus dilakukan karena grup ini menunjukkan aktivitas komunitas yang intens dan tema bervariasi dalam setiap cuitan. Pemilihan grup ini juga mempertimbangkan relevansi topik, ketersediaan data, dan potensi kontribusi praktis penelitian terhadap pemahaman pola interaksi pengguna di media sosial.

Penelitian ini memilih klasifikasi multikelas karena karakteristik data cuitan terkait grup K-pop aespa, keragaman topik dan kategori konten. Berbeda dengan klasifikasi biner atau berbasis sentimen, pendekatan multikelas memungkinkan setiap cuitan dikategorikan ke dalam kelas yang sesuai secara spesifik, sehingga analisis topik dan kategori cuitan dapat dilakukan secara lebih komprehensif. Pendekatan ini penting untuk memetakan keragaman opini, topik, dan interaksi pengguna di media sosial secara sistematis, sekaligus meminimalkan kehilangan informasi, bila data hanya diklasifikasikan dalam dua kategori saja. Pendekatan ini relevan karena berbagai penelitian sebelumnya telah menunjukkan bahwa klasifikasi multikelas mampu menangani variasi data kompleks secara efektif, sehingga menjadi relevan untuk penelitian klasifikasi cuitan X terkait grup aespa.

Representasi numerik diperlukan dalam konteks penelitian klasifikasi teks agar data teks mentah dapat diproses oleh algoritma *machine learning*. *Term Frequency-Inverse Document Frequency* (TF-IDF) dipilih karena kemampuannya memberikan bobot pada kata-kata yang lebih informatif dan membedakan kontribusi setiap kata terhadap klasifikasi. Manning et al. (2009) menjelaskan bahwa TF-IDF menghitung bobot kata berdasarkan frekuensi kemunculannya dalam dokumen serta seberapa umum kata tersebut di seluruh korpus (kumpulan dokumen), sehingga kata yang lebih unik atau spesifik terhadap topik memiliki pengaruh lebih besar terhadap model. Selain itu, TF-IDF efektif dalam menangani data teks berdimensi tinggi dan *sparse*, yang sering ditemukan pada cuitan, serta lebih sederhana dan interpretatif dibandingkan representasi lain seperti *word embeddings*, sehingga cocok untuk penerapan klasifikasi multikelas pada dataset teks di platform X.

Algoritma *Decision Tree* dan *Random Forest* dipilih sebagai metode klasifikasi karena keduanya merupakan algoritma serumpun dan menawarkan keunggulan untuk karakteristik data teks tidak terstruktur. *Decision Tree* memiliki interpretabilitas tinggi dan konstruksi keputusan yang relatif cepat, sehingga memudahkan pemahaman struktur klasifikasi dan interpretasi hasil. Namun, algoritma ini rentan terhadap *overfitting*, terutama pada data berdimensi tinggi dan *sparse*. Keterbatasan tersebut dapat diatasi oleh *Random Forest* sebagai algoritma *ensemble* berbasis pohon yang menggabungkan hasil *voting* dari banyak pohon keputusan, sehingga meningkatkan akurasi, stabilitas model, dan ketahanan terhadap *overfitting* (Breiman, 2001; Han et al., 2012). Selain itu, kedua algoritma ini telah banyak digunakan dalam penelitian klasifikasi teks, sehingga memungkinkan perbandingan kinerja yang relevan dengan studi sebelumnya.

Firnanda et al. (2025) membandingkan algoritma *Decision Tree* dan *Random Forest* dalam klasifikasi penjualan produk pada data Supermarket ASDA, dan menunjukkan bahwa *Random Forest* memiliki performa yang lebih baik dan lebih stabil dibandingkan *Decision Tree*, termasuk setelah dilakukan *hyperparameter tuning*. Oktavianto et al. (2024) membandingkan *Decision Tree*

dan *Random Forest* dalam klasifikasi teks data kesehatan, di mana *Random Forest* menunjukkan akurasi lebih tinggi dan stabil dibandingkan *Decision Tree*. Abdurrazik & Wirawan (2025) mengklasifikasikan teks cuitan (tweet) berbahasa Indonesia berdasarkan topik sosial menggunakan *Support Vector Machine* (SVM) dan representasi TF-IDF, serta menunjukkan bahwa metode tersebut mampu menghasilkan akurasi yang baik dalam klasifikasi teks pendek media sosial. Bintang et al. (2025), serta Chen et al. (2020) menegaskan efektivitas *Decision Tree* dan *Random Forest* dalam berbagai domain, termasuk prediksi fitur kritis dan seleksi fitur, menunjukkan bahwa kedua algoritma ini mampu menangani data berdimensi tinggi dan kompleks.

Berdasarkan literatur dan penelitian sebelumnya, masih diperlukan evaluasi langsung terhadap performa algoritma *Decision Tree* dan *Random Forest* dalam klasifikasi multikelas cuitan X terkait grup K-pop aespa. Hal ini penting untuk memahami bagaimana kedua algoritma menangani data teks media sosial yang dinamis, *sparse*, dan berdimensi tinggi, dengan ragam topik yang kompleks. Kesenjangan ini relevan untuk diteliti karena karakteristik cuitan X terkait grup aespa berbeda dari dataset teks pada penelitian sebelumnya, baik dari segi variasi topik maupun penggunaan bahasa informal.

Penelitian ini dilakukan untuk menguji dan membandingkan performa *Decision Tree* dan *Random Forest* dalam mengklasifikasikan cuitan terkait grup K-pop aespa menggunakan representasi *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagai dasar pengembangan metode klasifikasi teks berbasis statistika dan pembelajaran mesin. Penelitian ini bertujuan memberikan pemahaman yang lebih komprehensif mengenai efektivitas kedua algoritma dalam menghadapi beragamnya data teks media sosial. Hasil penelitian diharapkan tidak hanya menjadi acuan dalam ketepatan pemilihan algoritma klasifikasi dan memberikan kontribusi metodologis serta praktis bagi penelitian serupa di masa mendatang. Akan tetapi, juga menawarkan wawasan bagi pihak yang tertarik menganalisis topik atau interaksi pengguna di media sosial, sehingga pendekatan klasifikasi multikelas dapat diterapkan secara lebih efektif.

1.2 Rumusan Masalah

Berdasarkan latar belakang, penelitian ini difokuskan pada evaluasi performa algoritma klasifikasi dalam konteks data teks tidak terstruktur dari media sosial. Secara lebih rinci, masalah dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana langkah-langkah algoritma *Decision Tree* dan *Random Forest*?
2. Bagaimana kinerja algoritma *Decision Tree* dalam mengklasifikasikan cuitan X terkait grup K-pop aespa menggunakan representasi TF-IDF?
3. Bagaimana kinerja algoritma *Random Forest* dalam mengklasifikasikan cuitan X terkait grup K-pop aespa menggunakan representasi TF-IDF?
4. Algoritma manakah, antara *Decision Tree* dan *Random Forest*, yang lebih optimal dalam mengklasifikasikan data teks berupa cuitan dari platform X terkait grup K-pop aespa?

Rumusan masalah ini menjadi panduan penelitian untuk mengevaluasi secara sistematis kelebihan, kelemahan, dan relevansi masing-masing algoritma dalam konteks klasifikasi teks pada media sosial.

1.3 Batasan Masalah

Agar penelitian ini lebih terfokus dan hasilnya dapat dianalisis secara mendalam, ditetapkan beberapa batasan masalah sebagai berikut:

1. Sumber Data: Penelitian ini menggunakan data teks berupa cuitan grup K-pop aespa di platform X. Data diambil menggunakan metode scraping, dan tidak mencakup platform media sosial lain.
2. Jenis Algoritma: Analisis performa difokuskan pada dua algoritma supervised learning, yaitu Decision Tree dan Random Forest, tanpa mengevaluasi algoritma lain seperti SVM, Naive Bayes, atau neural network.

3. Representasi Teks: Semua dokumen teks diproses dan direpresentasikan dalam bentuk numerik menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), sehingga metode lain seperti word embedding tidak digunakan dalam penelitian ini.
4. Jenis Klasifikasi: Penelitian terbatas pada klasifikasi multikelas, di mana setiap cuitan dari platform X dikategorikan ke dalam kelas Promosi, Informasi, Opini/Ekspresi, dan Interaksi Fandom.
5. Evaluasi Model: Kinerja algoritma diukur menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, dengan pendekatan *macro-average* dan *weighted-average*.
6. *Preprocessing* Teks: Tahapan *preprocessing* mencakup *cleaning*, *case folding*, *tokenization*, *stopword removal*, dan *stemming/lemmatization*. Tahap lanjutan seperti *feature engineering* tambahan tidak menjadi fokus utama.

Batasan masalah ini ditetapkan untuk menjaga fokus penelitian, memudahkan pengolahan data, dan memastikan interpretasi hasil klasifikasi dapat dilakukan secara valid dan sistematis.

1.4 Tujuan Penelitian

Penelitian ini memiliki tujuan utama untuk mengevaluasi dan membandingkan performa algoritma klasifikasi dalam pengolahan data teks tidak terstruktur dari media sosial. Adapun tujuan spesifik dari penelitian ini adalah sebagai berikut:

1. Menjelaskan langkah-langkah algoritma *Decision Tree* dan *Random Forest*.
2. Menganalisis kinerja algoritma *Decision Tree* dalam mengklasifikasikan cuitan X terkait grup K-pop aespa menggunakan representasi TF-IDF.

3. Menganalisis kinerja algoritma *Random Forest* dalam mengklasifikasikan cuitan terkait grup K-pop aespa menggunakan representasi TF-IDF, serta mengevaluasi stabilitas dan akurasi prediksi model.
4. Membandingkan kinerja *Decision Tree* dan *Random Forest* dalam klasifikasi data teks berupa cuitan dari platform X berdasarkan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk konteks multikelas.

Tujuan penelitian ini diharapkan memberikan kontribusi metodologis dalam pemilihan algoritma klasifikasi teks, serta memberikan wawasan praktis bagi pengolahan data media sosial tidak terstruktur dan berdimensi tinggi.

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat baik secara teoritis maupun praktis. Manfaat tersebut dapat dijabarkan sebagai berikut:

Manfaat Teoritis:

1. Memberikan kontribusi terhadap pengembangan ilmu pengetahuan dalam bidang *text mining*, khususnya terkait pemrosesan dan klasifikasi data teks tidak terstruktur menggunakan algoritma *Decision Tree* dan *Random Forest*.
2. Menjadi referensi metodologis bagi peneliti lain yang ingin mengevaluasi atau membandingkan performa algoritma klasifikasi pada data media sosial, termasuk aspek representasi TF-IDF dan penanganan data *high-dimensional* serta *sparse*.

Manfaat Praktis:

1. Memberikan panduan bagi praktisi atau analis data dalam memilih algoritma klasifikasi yang efektif untuk pengolahan data teks di platform X, khususnya dalam konteks klasifikasi multikelas.

2. Membantu organisasi, agensi, atau pengelola media sosial memahami tren konten terkait topik tertentu (dalam penelitian ini: grup K-pop aespa) melalui pemanfaatan model klasifikasi berbasis *machine learning*.
3. Menjadi dasar pengembangan sistem otomatisasi analisis teks, misalnya dalam *monitoring* media sosial atau pengambilan keputusan berbasis data, dengan pertimbangan efisiensi dan akurasi model.

Hasil penelitian ini diharapkan memperkuat integrasi antara teori dan praktik dalam pemrosesan data teks, serta memberikan kontribusi nyata bagi pemanfaatan data media sosial secara sistematis. Selain itu, penelitian ini dapat menjadi referensi bagi studi klasifikasi multikelas di platform media sosial lain, sehingga pendekatan dan temuan yang diperoleh bersifat lebih luas dan aplikatif.

1.6 Tinjauan Pustaka

Beberapa penelitian terdahulu telah mengevaluasi algoritma klasifikasi multikelas, terutama dalam konteks data teks, data berdimensi tinggi, atau data media sosial, dan dapat menjadi acuan bagi penelitian ini. Referensi yang dipilih saling melengkapi karena memberikan gambaran tentang (1) teknik representasi data teks, (2) pemilihan dan optimasi dataset, (3) metode klasifikasi multikelas, serta (4) evaluasi performa algoritma *Decision Tree* dan *Random Forest*.

Firnanda et al. (2025) melakukan analisis perbandingan algoritma *Decision Tree* dan *Random Forest* dalam klasifikasi penjualan produk pada data Supermarket ASDA. Penelitian tersebut menunjukkan bahwa *Random Forest* memiliki performa yang lebih baik dibandingkan *Decision Tree*, baik pada pemodelan dasar maupun setelah dilakukan *hyperparameter tuning*, dengan tingkat akurasi mencapai 99%. Hasil evaluasi menggunakan *precision*, *recall*, dan *F1-score* juga mengindikasikan bahwa *Random Forest* lebih stabil dan akurat. Meskipun konteks data yang digunakan bukan data teks, penelitian ini memperkuat temuan bahwa *Random Forest* secara konsisten mampu mengatasi kelemahan *Decision Tree*, khususnya terkait *overfitting*.

Oktavianto et al. (2024) meneliti klasifikasi teks data kesehatan menggunakan *Decision Tree* dan *Random Forest*. Penelitian ini menggunakan empat skenario pembagian data latih dan data uji (10%-90%, 15%-85%, 20%-80%, 25%-75%). Hasil penelitian menunjukkan bahwa *Random Forest* selalu memiliki akurasi lebih tinggi dibanding *Decision Tree*, dengan akurasi stabil pada nilai 99%, sedangkan *Decision Tree* berada di kisaran 75%. Penelitian ini menekankan keunggulan *Random Forest* dalam menangani dataset multikelas yang kompleks serta memberikan stabilitas prediksi yang lebih tinggi.

Bintang et al. (2025) membandingkan lima algoritma klasifikasi yaitu *Logistic Regression*, *Naive Bayes*, *Decision Tree*, *Random Forest*, dan *SVM*. Klasifikasi ini digunakan untuk analisis sentimen ulasan pengguna aplikasi Netflix dari *Google Play Store*. Data diproses melalui pembersihan teks, tokenisasi, *stopword removal*, dan *stemming*, serta direpresentasikan menggunakan TF-IDF. Evaluasi dilakukan berdasarkan rasio 90:10 antara data latih dan data uji. Hasil penelitian menunjukkan bahwa *Logistic Regression* dan *Random Forest* memiliki akurasi tertinggi sebesar 76%, diikuti *SVM* 74%, *Decision Tree* 73%, dan *Naive Bayes* 71%. Penelitian ini relevan karena menggunakan TF-IDF dalam klasifikasi multikelas dan membandingkan algoritma populer pada data teks nyata.

Chen et al. (2020) menekankan pentingnya seleksi fitur pada dataset berdimensi tinggi untuk meningkatkan performa klasifikasi. Penelitian ini menggunakan tiga dataset populer, yaitu *Bank Marketing*, *Car Evaluation*, dan *Human Activity Recognition*. *Random Forest* efektif dalam memilih fitur penting dan meningkatkan akurasi klasifikasi. Hasil eksperimen menunjukkan akurasi tertinggi *Random Forest*, misalnya 98,57% pada 561 fitur dan 93,26% pada 6 fitur. Penelitian ini relevan karena menekankan keunggulan *Random Forest* dalam menangani data sparse dan high-dimensional, serta mengombinasikannya dengan *feature selection* untuk meningkatkan kinerja klasifikasi.

Penelitian lain dalam konteks klasifikasi teks media sosial dilakukan oleh Abdurrazik & Wirawan (2025) yang mengklasifikasikan tweet berbahasa Indonesia ke dalam tiga kategori topik, yaitu politik, hiburan, dan lainnya menggunakan

algoritma *Support Vector Machine* (SVM) dengan representasi TF-IDF. Tahapan *preprocessing* meliputi pembersihan teks, *case folding*, tokenisasi, *stopword removal*, dan *stemming*. Hasil penelitian menunjukkan bahwa model mampu mencapai akurasi sebesar 84%, yang menunjukkan bahwa kombinasi TF-IDF dan algoritma klasifikasi efektif dalam menangani data teks pendek dari media sosial. Penelitian ini memiliki kesamaan pada jenis data penelitian, yaitu teks cuitan (tweet), serta tahapan *preprocessing* dan representasi fitur. Namun demikian, penelitian tersebut berfokus pada klasifikasi berdasarkan topik, sedangkan penelitian ini mengklasifikasikan berdasarkan tujuan komunikasi, sehingga memiliki tingkat kompleksitas berbeda.

Dengan kombinasi beberapa studi berikut ini, penelitian dapat membangun kerangka metodologis yang kuat dan relevan untuk klasifikasi multikelas cuitan di platform media sosial X.

Tabel 1. 1 Tinjauan pustaka

No	Judul	Hasil	Persamaan	Perbedaan
1.	Analisis Perbandingan *Decision Tree* dan *Random Forest* dalam Klasifikasi Penjualan Produk pada Supermarket (Firnanda et al., 2025)	Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memiliki performa yang lebih baik dibandingkan *Decision Tree*, baik pada pemodelan dasar maupun setelah dilakukan *hyperparameter tuning*, dengan tingkat akurasi mencapai 99%. Nilai	Membandingkan algoritma *Decision Tree* dan *Random Forest* serta menggunakan metrik evaluasi akurasi, *precision*, *recall*, dan *F1-score*.	Menggunakan dataset penjualan supermarket untuk skema klasifikasi biner (produk laris dan tidak laris) dengan menggunakan variabel numerik dan kategorikal.

		precision, *recall*, dan *F1-score* pada *Random Forest* juga lebih tinggi dibandingkan *Decision Tree*.		
2.	Analisis Komparasi Kinerja Metode *Decision Tree* dan *Random Forest* dalam Klasifikasi Teks Data Kesehatan (Oktavianto et al., 2024)	Penelitian ini membandingkan *Decision Tree* dan *Random Forest* untuk klasifikasi teks data kesehatan dengan pembagian data latih dan uji 10%-90%, 15%-85%, 20%-80%, dan 25%-75%. Hasil penelitian menunjukkan bahwa *Random Forest* memiliki akurasi lebih tinggi dan stabil (~99%) dibanding *Decision Tree* (~75%).	Menggunakan *Decision Tree* dan *Random Forest* serta melakukan evaluasi performa model klasifikasi.	Dataset berupa teks kesehatan, bukan data media sosial terkait K-pop; analisis dilakukan berdasarkan beberapa skenario pembagian data, bukan topik K-pop aespa; fokus pada konteks medis, bukan media sosial.
3.	Perbandingan Kinerja Algoritma Klasifikasi pada Review Pengguna	Penelitian ini menganalisis klasifikasi sentimen ulasan pengguna aplikasi Netflix dari Google Play Store.	Menggunakan representasi TF-IDF untuk klasifikasi teks multikelas, serta	Dataset berupa review Netflix, bukan data teks media sosial terkait K-pop; fokus

	Aplikasi Netflix (Bintang et al., 2025)	Data diproses melalui pembersihan teks, tokenisasi, *stopword removal*, dan *stemming*, serta direpresentasikan menggunakan TF-IDF. Lima algoritma dibandingkan: *Logistic Regression*, *Naive Bayes*, *Decision Tree*, *Random Forest*, dan *SVM*. Hasil penelitian menunjukkan bahwa *Random Forest* dan *Logistic Regression* memiliki akurasi tertinggi sebesar 76%, diikuti *SVM* 74%, *Decision Tree* 73%, dan *Naive Bayes* 71%.	membandingkan performa beberapa algoritma, termasuk *Decision Tree* dan *Random Forest*.	penelitian pada analisis sentimen, bukan klasifikasi topik; pembagian data dan jumlah algoritma yang diuji berbeda.

4.	*Selecting Critical Features for Data Classification Based on Machine Learning Methods* (Chen et al., 2020)	Penelitian ini menekankan pentingnya seleksi fitur untuk dataset berdimensi tinggi, menggunakan tiga dataset, yaitu *Bank Marketing*, *Car Evaluation*, dan *Human Activity Recognition*. *Random Forest* terbukti efektif dalam memilih fitur penting, mengurangi dimensi, serta meningkatkan akurasi dan stabilitas klasifikasi. Hasil eksperimen menunjukkan akurasi tertinggi *Random Forest*: 98,57% pada 561 fitur, dan 93,26% pada 6 fitur.	Menggunakan *Random Forest* untuk klasifikasi data berdimensi tinggi (*high-dimensional*) dan menguji performa algoritma dalam konteks multikelas.	Tidak menggunakan algoritma *Decision Tree* dan representasi TF-IDF; dataset berbeda, bukan data media sosial.

5.	Analisis Klasifikasi Tweet Berdasarkan Topik Sosial Menggunakan SVM (Abdurrazik & Wirawan, 2025)	Model klasifikasi menggunakan SVM dengan representasi TF-IDF mampu mencapai akurasi sekitar 84%, menunjukkan efektivitas metode dalam mengklasifikasikan cuitan atau teks pendek dari platform media sosial X.	Menggunakan data teks dari platform media sosial X (sebelumnya Twitter), melalui tahapan *preprocessing* serta representasi TF-IDF dalam klasifikasi multikelas.	Menggunakan algoritma *Support Vector Machine* (SVM) dan klasifikasi berbasis topik (politik, hiburan, lainnya), sedangkan penelitian ini menggunakan *Decision Tree* dan *Random Forest* serta klasifikasi berbasis konten (informasi, opini, promosi, interaksi).

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini terdiri dari enam bab, di antaranya yaitu bab pendahuluan, landasan teori, metode penelitian, pembahasan, studi kasus, dan penutup. Adapun uraian setiap bab adalah sebagai berikut:

1. BAB I PENDAHULUAN

   Bab ini membahas latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, tinjauan pustaka, dan sistematika penulisan.

2. BAB II LANDASAN TEORI

   Bab ini berisi teori-teori yang menjadi dasar penelitian, mencakup konsep dasar matematika, konsep data, hingga konsep dasar yang menjadi penunjang algoritma pada penelitian ini.

3. BAB III METODE PENELITIAN

   Bab ini menjelaskan jenis penelitian, metode pengumpulan data, variabel penelitian, metode penelitian, metode analisis data, serta flowchart penelitian.

4. BAB IV PEMBAHASAN

   Bab ini membahas mengenai algoritma yang digunakan pada penelitian, yaitu algoritma *Decision Tree* dan *Random Forest*, pembahasan mengenai evaluasi model, serta contoh sederhana penerapan kedua algoritma.

5. BAB V STUDI KASUS

   Bab ini membahas pengaplikasian algoritma *Decision Tree* dan *Random Forest* pada klasifikasi multikelas cuitan X terkait grup K-pop aespa.

6. BAB VI PENUTUP

   Bab ini berisi kesimpulan dari hasil penelitian yang dilakukan dan saran untuk penelitian atau pengembangan selanjutnya.

BAB VI
PENUTUP

6.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan pada studi kasus, berikut adalah kesimpulan dari analisis kinerja algoritma *Decision Tree* dan *Random Forest* pada klasifikasi multikelas cuitan X terkait grup K-pop aespa menggunakan TF-IDF:

1. Penelitian ini telah menerapkan proses klasifikasi multikelas pada data teks tidak terstruktur berupa cuitan platform X terkait grup K-pop aespa. Dataset penelitian diperoleh melalui proses *scraping* dan *sampling* data, kemudian dilakukan pelabelan manual dengan satu annotator yaitu peneliti. Label terdiri dari empat kategori, yaitu Informasi, Opini/Ekspresi, Interaksi Fandom, dan Promosi. Selanjutnya, data diproses melalui tahapan *preprocessing* yang meliputi *case folding*, *cleaning*, *tokenization*, *stopword removal*, dan *stemming* untuk mengurangi noise pada teks. Setelah itu, representasi fitur dilakukan menggunakan metode TF-IDF sehingga data teks dapat diubah menjadi representasi numerik untuk digunakan dalam proses klasifikasi *machine learning*.
2. Algoritma *Decision Tree* mampu melakukan klasifikasi multikelas terhadap data cuitan terkait aespa dengan nilai akurasi sebesar 70% dan *weighted F1-score* sebesar 0.70. Proses klasifikasi pada *Decision Tree* dilakukan dengan membentuk struktur pohon keputusan berdasarkan fitur-fitur TF-IDF yang memberikan pemisahan kelas terbaik pada setiap *node*. Berdasarkan hasil evaluasi menggunakan *confusion matrix* dan *classification report*, model menunjukkan performa yang cukup baik pada kategori Promosi dan Informasi. Namun, model masih mengalami kesulitan dalam membedakan kategori Interaksi Fandom.

3. Algoritma *Random Forest* menghasilkan performa klasifikasi yang lebih baik dibandingkan *Decision Tree* dengan nilai akurasi sebesar 75% dan *weighted F1-score* sebesar 0.74. Model *Random Forest* dibangun menggunakan parameter $\text{n\_estimators} = 100$ yang membentuk 100 pohon keputusan dalam proses klasifikasi. Meskipun demikian, kategori Interaksi Fandom masih menjadi kategori dengan performa terendah karena karakteristik bahasa pada kategori tersebut masih memiliki kemiripan dengan kategori Opini/Ekspresi.
4. Berdasarkan perbandingan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, *Random Forest* menunjukkan performa klasifikasi yang lebih baik dibandingkan *Decision Tree* pada dataset penelitian ini. Berdasarkan hasil evaluasi, *Random Forest* menunjukkan peningkatan performa pada sebagian besar kategori, terutama Opini/Ekspresi dan Promosi. Dengan demikian, hasil penelitian menunjukkan bahwa pendekatan *ensemble* pada *Random Forest* lebih efektif dalam menangani klasifikasi multikelas data teks media sosial dibandingkan pohon keputusan tunggal. Namun, distribusi data yang tidak sepenuhnya seimbang serta kemiripan karakteristik bahasa antar kategori masih menjadi tantangan dalam proses klasifikasi cuitan terkait grup aespa atau cuitan fandom pada platform media sosial X.

6.2 Saran

Penulis menyadari bahwa penelitian ini masih memiliki beberapa keterbatasan dalam proses analisis kinerja algoritma *Decision Tree* dan *Random Forest* pada klasifikasi multikelas cuitan X terkait grup K-pop aespa menggunakan TF-IDF. Oleh karena itu, berikut beberapa saran untuk pengembangan pada penelitian selanjutnya:

1. Penggunaan jumlah data yang lebih besar serta periode pengambilan data yang lebih panjang dapat membantu menghasilkan variasi bahasa dan konteks cuitan yang lebih beragam. Selain itu, perbandingan dengan algoritma lain seperti *Support Vector Machine* atau *XGBoost*, maupun

metode *deep learning* seperti LSTM atau BERT dapat dilakukan untuk memperoleh gambaran performa klasifikasi yang lebih luas pada data teks media sosial.

2. Penerapan teknik penanganan data tidak seimbang, seperti SMOTE, *undersampling*, atau *oversampling*, dapat dipertimbangkan untuk meningkatkan performa klasifikasi pada kelas minoritas.
3. Proses pelabelan data dapat dikembangkan dengan melibatkan lebih dari satu annotator sehingga konsistensi dan reliabilitas hasil pelabelan menjadi lebih baik. Penggunaan pengukuran *inter-annotator agreement* juga dapat dipertimbangkan untuk mengurangi subjektivitas dalam proses anotasi data.

DAFTAR PUSTAKA

Abdurrazik, & Wirawan, I. M. W. (2025). Analisis Klasifikasi Tweet Berdasarkan Topik Sosial Menggunakan SVM. *Jurnal Nasional Teknologi Informasi Dan Aplikasinya*, *3*(4), 855–864.

Aggarwal, C. C., & Zhai, C. X. (2012). *Mining text data*. Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-3223-4

Alyasiri, O. M., & Cheah, Y. N. (2025). Multi-Class Text Classification using Machine Learning Techniques. *Engineering, Technology and Applied Science Research*, *15*(3), 22598–22604. https://doi.org/10.48084/etasr.9994

Amaliana, L., Astuti, A. B., Gadis, R. S., Rabbani, N. A., & Sevia, N. A. (2025). *Prediksi Resiko Kematian Penderita Gagal Ginjal Kronis Dengan Voting Classifier Dan Random Forest Pada Data Tidak Seimbang*. *12*(4), 859–866.

Anton, H., & Kaul, A. (2019). *Elementary Linear Agebra* (12th ed.). John Wiley & Sons.

Anton, H., & Rorres, C. (2010). *Elementary Linear Algebra: Applications Version*. John Wiley & Sons.

Bintang, R. A. K. N., Romadloni, N. T., & Ramadhan, F. (2025). Perbandingan Kinerja Algoritma Klasifikasi Pada Review Pengguna Aplikasi Netflix. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(2). https://doi.org/10.23960/jitet.v13i2.6303

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science+Business Media.

Breiman, L. (2001). Random Forests. In *Machine Learning* (Vol. 45). Kluwer Academic Publishers.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal*

of Big Data, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

De Boe, B. (2014). Use Cases for Unstructured Data. *Intersystems White Paper*.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, *1857*, 1–15. https://doi.org/10.1007/3-540-45014-9

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*, 26. https://doi.org/10.1155/2022/1883698

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50. https://doi.org/10.14445/22312803/ijctt-v38p109

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook Text: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Firnanda, P. A., Shofwatillah, L., Rahma, F., & Fauzi, F. (2025). Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket. *Emerging Statistics and Data Science Journal*, *3*(1), 445–461. https://doi.org/10.20885/esds.vol3.iss.1.art2

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. In *The Mathematical Intelligencer* (2nd ed.). Springer.

Imrona, M. (2013). *Aljabar Linear Dasar*. Erlangga.

Işik, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(3), 1405–1421. https://doi.org/10.3906/elk-1907-46

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Science+Business Media.

https://doi.org/10.1007/978-1-4614-7138-7

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(1), 20–28. https://doi.org/10.38094/jastt20165

Jin, D. Y. (2016). *New Korean Wave: Transnational Cultural Power in the Age of Social Media.* University of Illinois Press. http://www.jstor.org/stable/10.5406/j.ctt18j8wkv

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Khairudin, K., Machfud, S., & Cahyono, Y. (2024). Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode C4.5. *KERNEL: Jurnal Riset Inovasi Bidang Informatika Dan Pendidikan Informatika*, *5*(2), 83–93. https://doi.org/10.31284/j.kernel.2024.v5i2.7315

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE.

Krüger, F. (2016). *Activity, Context, and Plan Recognition with Computational Causal Behaviour Models*. University of Rostock.

Kurniadi, D., Pertiwi, A. I., & Mulyani, A. (2025). Ensemble Voting Classifier Berbasis Multi-Algoritma dan Metode SMOTE untuk Klasifikasi Penyakit Jantung. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *14*(2), 145–153. https://doi.org/10.22146/jnteti.v14i2.17157

Lay, D. C., Lay, S. R., & McDonald, J. J. (2022). *Linear Algebra and Its*

Applications. Pearson Education Limited.

Majeed, N. M., & Ramo, F. M. (2022). Performance Evaluation of the Ensemble and Selected Machine Learning Techniques. *Journal of Modern Computing and Engineering Research*, *2022*, 94–100.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1108/00242530410565256

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python*. O'Reilly Media, Inc.

Nasrullah, R. (2021). *Media Sosial Persfektif Komunikasi, Budaya, dan Sosioteknologi*. PT Remaja Rosdakarya.

Nasution, M., Munthe, I. R., Nasution, F. A., & Defit, S. (2025). Optimizing Text Classification Using Techniques AdaBoost Ensemble with Decision Tree Algorithm. *CogITo Smart Journal*, *11*(1), 39–51. https://doi.org/10.31154/cogito.v11i1.741.39-51

Oktavianto, H., Sulistyo, H. W., Wijaya, G., Irawan, D., & Abdurrahman, G. (2024). Analisis Perbandingan Decision Tree dan Random Forest Pada Klasifikasi Teks Data Kesehatan. *Bina Insani Ict Journal*, *11*(1), 56.

Permana, A. A., S, W., Santoso, L. W., Wibowo, G. W. N., Wardhani, A. K., Rahmaddeni, Wahidin, A. J., Yuliastuti, G. E., Elisawati, Wijayanti, R. R., & Abdurrasyid. (2023). *Machine learning*.

Quinlan, J. R. (1986). Induction of Decision Trees J.R. *Machine Learning*, *1*(1), 81–106.

Seni, G., & Elder, J. F. (2010). *Ensemble Methods in Data Mining: Improving AccuracyThrough Combining Predictions*. Morgan & Claypool.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System*

Technical Journal, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sint, R., Schaffert, S., Stroka, S., & Ferstl, R. (2009). Combining unstructured, fully structured and semi-structured information in semantic wikis. *CEUR Workshop Proceedings*, *464*, 73–87.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

Zainudin, M. N. S., Sulaiman, M. N., Musapha, N., Perumal, T., & Mohamed, R. (2018). Solving Classification Problem Using Ensemble Binarization Classifier. *International Journal of Engineering & Technology*, *7*(4.31), 280–284.

Zhang, M. L., & Zhang, K. (2010). Multi-label learning by exploiting label dependency. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 999–1007. https://doi.org/10.1145/1835804.1835930

Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. Mc Graw-Hill Companies.