DIGINTEL-AI: DIGITAL INNOVATION AND INTELLIGENCE – AI
e-ISSN: 3123-8076
Journal homepage: https://journal.ajirapublisher.id/index.php/digintelai

Klasifikasi Konten AI-*Generated* dan Non-AI pada Teks *Hashtag* dan Komentar TikTok Menggunakan *Logistic Regression*

Classification of AI-Generated and Non-AI Content Using TikTok Hashtag and Comment Text with Logistic Regression

Khalifah Safa Rizky[1], Irwansyah[2]*, Tupan Tri M[3]

[1,2]Teknik Informatika, Fakultas Teknologi Industri dan Informatika, Universitas Muhammadiyah Prof. Dr. HAMKA, Jakarta, Indonesia
[3]Sistem Informasi, Fakultas Teknologi, Institut Sosial dan Teknologi Widuri, Jakarta, Indonesia

*Corresponding Author: irwansyah@uhamka.ac.id

ARTICLE INFO

Vol. 2, No. 1, October 2026, pp. 1-10.
DOI: https://doi.org/10.66217/digintel-ai.v2i1.21

Article history:
Revised: 15 October 2026
Accepted: 20 Octobet 2026
Published: 31 October 2026

Copyright © 2026 by the authors. Licensee DIGINTEL-AI, Jakarta, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License (CC BY-NC-SA 4.0) (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Abstract

The widespread distribution of AI-generated content on TikTok raises concerns about disinformation and the authenticity of online interaction. Most previous studies emphasise visual features, whereas textual features such as hashtags and user comments remain underexplored. This study aims to classify AI-generated and non-AI content on TikTok using the text of hashtags and user comments with the Logistic Regression method. Data were collected through web scraping with Apify and, after cleaning, 1,008 text records were obtained. Preprocessing consisted of case folding, removal of URLs, mentions, hashtags and non-alphabetic characters, tokenisation, stop-word filtering and stemming. The text was converted into numerical features using Term Frequency–Inverse Document Frequency (TF-IDF), and the data were split 70:30 into 705 training and 303 testing records. The model was built and evaluated in RapidMiner Studio using a confusion matrix. It achieved an overall accuracy of 72.28%. For the non-AI class, precision, recall and F1-score were 74.45%, 86.67% and 80.09%, whereas for the AI-generated class they were 65.79%, 46.30% and 54.35%. The low recall of the AI-generated class was caused by 58 false negatives. These findings indicate that Logistic Regression with TF-IDF can serve as a lightweight baseline for text-based detection of AI-generated content on TikTok, but class imbalance and the ambiguity of short, informal text still limit its sensitivity.

Keywords: *AI-Generated Content; Logistic Regression; TF-IDF; Text Classification; TikTok.*

Abstrak

Maraknya penyebaran konten AI-*generated* di TikTok menimbulkan kekhawatiran terhadap disinformasi dan keaslian interaksi daring. Sebagian besar penelitian terdahulu menitikberatkan pada fitur visual, sedangkan fitur teks seperti *hashtag* dan komentar pengguna belum banyak dikaji. Penelitian ini bertujuan mengklasifikasikan konten AI-*generated* dan non-AI di TikTok berdasarkan teks *hashtag* dan komentar pengguna menggunakan metode *Logistic Regression*. Data dikumpulkan melalui *web scraping* dengan Apify dan, setelah proses *cleaning*, diperoleh 1.008 data teks. Pra-pemrosesan meliputi *case folding*, penghapusan URL, *mention*, *hashtag*, dan karakter non-alfabet, tokenisasi, penyaringan *stopword*, dan *stemming*. Teks diubah menjadi

fitur numerik dengan pembobotan *Term Frequency–Inverse Document Frequency* (TF-IDF), kemudian data dibagi 70:30 menjadi 705 data latih dan 303 data uji. Model dibangun dan dievaluasi menggunakan RapidMiner Studio dengan *confusion matrix*. Model memperoleh akurasi keseluruhan 72,28%. Pada kelas non-AI, presisi, *recall*, dan *F1-score* masing-masing 74,45%, 86,67%, dan 80,09%, sedangkan pada kelas AI-*generated* masing-masing 65,79%, 46,30%, dan 54,35%. Rendahnya *recall* kelas AI-*generated* disebabkan oleh 58 kasus *false negative*. Hasil ini menunjukkan bahwa *Logistic Regression* dengan TF-IDF dapat menjadi *baseline* yang ringan untuk deteksi konten AI-*generated* berbasis teks di TikTok, namun ketidakseimbangan kelas dan ambiguitas teks yang pendek dan informal masih membatasi sensitivitasnya.

Kata Kunci: Konten AI-*Generated*; Klasifikasi Teks; *Logistic Regression*; TF-IDF; TikTok.

1. Pendahuluan

Transformasi digital telah mengubah cara masyarakat mengakses dan menyebarkan informasi. Media sosial kini menjadi sarana utama untuk berbagi konten digital, baik berupa teks, gambar, maupun video. TikTok merupakan salah satu platform yang paling populer, termasuk di Indonesia, khususnya di kalangan anak muda. Di balik popularitasnya, muncul tantangan baru berupa meningkatnya konten digital yang keasliannya diragukan, dan TikTok telah teridentifikasi sebagai salah satu wahana penyebaran disinformasi (Alonso-López *et al.*, 2021). Penyalahgunaan teknologi sintetis untuk memanipulasi informasi pun dinilai berdampak serius terhadap kepercayaan publik (Fadhilah, 2023).

Tantangan keaslian ini semakin besar seiring pesatnya perkembangan AI generatif. Model berbasis *transformer* seperti BERT (Devlin *et al.*, 2019) dan *large language models* (LLM) kini mampu menghasilkan teks yang fasih, relevan, dan terstruktur dengan baik secara tata bahasa. Akibatnya, komentar dan teks buatan mesin sulit dibedakan dari pendapat nyata pengguna. Jika teks dan *hashtag* yang dihasilkan AI mendominasi ruang digital, pendapat alami manusia dapat tertutup, kepercayaan terhadap platform menurun, dan analisis tren menjadi menyesatkan (Putri *et al.*, 2026). Di sisi lain, konten berbasis AI maupun non-AI di TikTok sama-sama dapat memengaruhi persepsi audiens, sehingga pembedaan keduanya penting untuk dipahami (Rahayu & Sudjradjat, 2025).

Berbagai penelitian telah memanfaatkan algoritma *machine learning* klasik untuk mengklasifikasikan teks media sosial. Rendra *et al.* (2024) melaporkan bahwa *Logistic Regression* dengan pembobotan TF-IDF menghasilkan akurasi tertinggi (80,41%) dibandingkan *Naïve Bayes* dan SVM pada komentar TikTok, dan Rizal (2024) memperoleh akurasi 81% pada klasifikasi sentimen komentar TikTok. Alexander *et al.* (2023) menggabungkan *text mining* dan *Logistic Regression* untuk menganalisis opini terhadap konten *deepfake*, sedangkan Cahyani *et al.* (2025) menunjukkan bahwa *Logistic Regression* tetap kompetitif pada data yang tidak seimbang. Pada ulasan aplikasi dan teks media sosial X, Agistina *et al.* (2026) memperoleh akurasi 79,78% dengan *Naïve Bayes* dan pembobotan TF-IDF, sedangkan Nida *et al.* (2026) melaporkan bahwa SVM (94,2%) mengungguli *Naïve Bayes* (80,5%). Pada ranah deteksi teks buatan mesin, Guo *et al.* (2023) dan Uchendu *et al.* (2021) membangun korpus pembanding dan menunjukkan bahwa detektor berbasis *deep learning* lebih tangguh, sementara Mitchell *et al.* (2023) mengusulkan pendekatan *zero-shot*. Ringkasan penelitian terkait disajikan pada Tabel 1.

Tabel 1. Ringkasan Penelitian Terkait

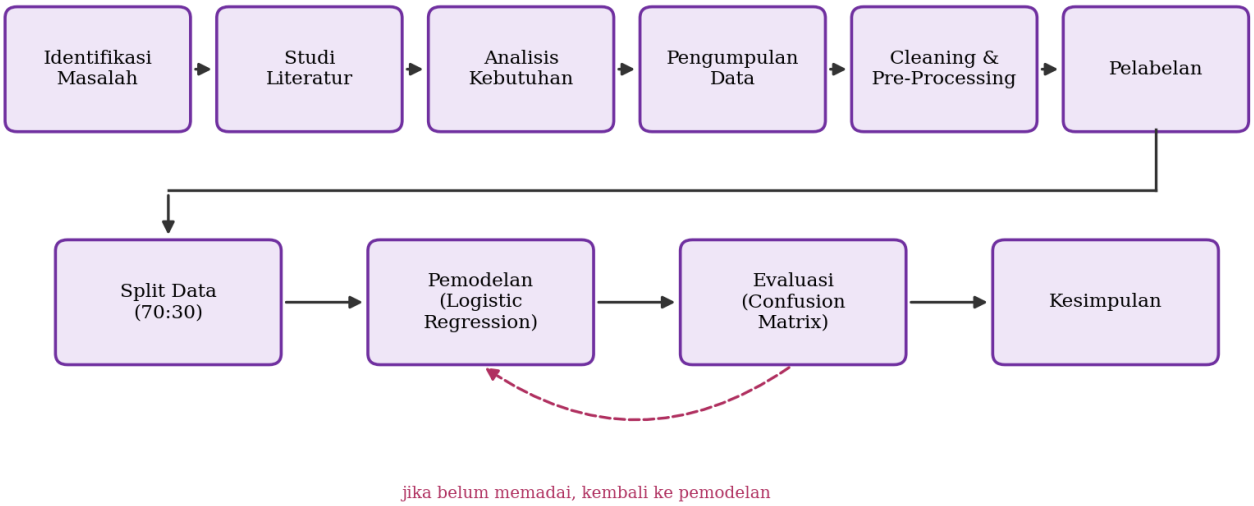
Peneliti	Metode dan data	Temuan utama
Rendra *et al.* (2024)	LR, *Naïve Bayes*, SVM; TF-IDF; komentar TikTok	LR akurasi tertinggi 80,41% (*Naïve Bayes* 79,33%; SVM 78,91%)
Rizal (2024)	LR; teks komentar TikTok	Akurasi 81% untuk klasifikasi sentimen
Alexander *et al.* (2023)	*Text mining* dan LR; opini tentang konten *deepfake*	Respons dominan negatif (60%); akurasi 75–85%
Cahyani *et al.* (2025)	LR dan *Naïve Bayes*; deteksi berita hoaks	LR mencapai *F1-score* 0,79 pada data tidak seimbang
April *et al.* (2025)	*Naïve Bayes* dan SVM; 5.000 komentar TikTok	SVM paling sesuai untuk teks pendek dan informal
Guo *et al.* (2023)	GLTR (fitur statistik) dan RoBERTa; korpus HC3	Detektor berbasis *deep learning* lebih tangguh
Wu *et al.* (2025)	Survei deteksi teks hasil LLM	Detektor rentan pada domain/bahasa baru; standar evaluasi belum seragam

Meskipun demikian, masih terdapat celah penelitian. Sebagian besar kajian konten sintetis di TikTok berfokus pada aspek visual, misalnya deteksi *deepfake* pada video (Abidin *et al.*, 2022), dan belum banyak yang mengkaji fitur teks seperti *hashtag* dan komentar pengguna, padahal kedua fitur tersebut dapat menjadi indikator untuk mengenali pola penyebaran konten manipulatif (Daulay *et al.*, 2025). Selain itu, kajian deteksi teks AI cenderung langsung beralih ke arsitektur *deep learning* yang membutuhkan sumber daya komputasi besar, padahal detektor yang ada pun rentan terhadap pergeseran domain dan manipulasi teks (Sadasivan *et al.*, 2023; Wu *et al.*, 2025). Algoritma klasik seperti *Logistic Regression* berpotensi menjadi alternatif yang ringan, efisien, serta memiliki koefisien yang dapat diinterpretasikan untuk teks berdimensi tinggi (Nirmala & Sanjaya, 2024). Dengan demikian, terdapat kesenjangan antara cepatnya perkembangan konten AI-*generated* di TikTok dan terbatasnya sistem klasifikasi berbasis teks yang sederhana.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk (1) menganalisis karakteristik konten AI-*generated* dan non-AI pada video pendek TikTok melalui *hashtag* dan komentar, (2) mengimplementasikan metode *Logistic Regression* dengan fitur TF-IDF untuk mengklasifikasikan kedua jenis konten tersebut, dan (3) mengevaluasi kinerja model menggunakan akurasi, presisi, *recall*, dan *F1-score*. Kontribusi penelitian ini adalah *baseline* klasifikasi berbasis teks yang ringan dan dibangun tanpa pemrograman (*no-code*) menggunakan RapidMiner, disertai analisis kesalahan pada kelas AI-*generated* yang dapat menjadi acuan bagi penelitian selanjutnya.

2. Metodologi

Penelitian ini dilakukan melalui tahapan yang ditunjukkan pada Gambar 1, yaitu identifikasi masalah, studi literatur, analisis kebutuhan, pengumpulan data, *cleaning* dan pra-pemrosesan data, pelabelan, pembagian data, pemodelan dengan *Logistic Regression*, evaluasi, dan penarikan kesimpulan. Apabila hasil evaluasi belum memadai, proses kembali ke tahap pemodelan.



Gambar 1. Alur Penelitian

1. Analisis Kebutuhan

Perangkat keras yang digunakan adalah prosesor AMD Ryzen 3 7320U dengan Radeon Graphics, RAM 8 GB, kartu grafis AMD Radeon Graphics, dan SSD NVMe 512 GB. Perangkat lunak yang digunakan adalah sistem operasi Windows 11 Home Single Language 64-bit, RapidMiner Studio versi 9.0.003 untuk pengolahan data dan pemodelan secara visual (*no-code*), serta Mendeley versi 2.125.2 untuk manajemen referensi. RapidMiner dipilih karena merupakan perangkat yang mudah diakses untuk membangun alur klasifikasi tanpa pemrograman (Setyawan *et al.*, 2025).

2. Pengumpulan Data

Data berupa teks *hashtag* dan komentar dari konten TikTok dikumpulkan menggunakan teknik *web scraping* melalui platform Apify. Data yang terkumpul terdiri atas 800 data *hashtag* dan 1.017 data komentar, disimpan dalam format Excel, kemudian diimpor ke RapidMiner. Data *hashtag* dan komentar digabungkan menjadi satu kesatuan dan, setelah proses *cleaning*, diperoleh 1.008 data teks. Setiap data memiliki tiga atribut utama: *Label* (AI-*Generated* atau Non-AI), *Source Type* (*hashtag* atau komentar pengguna), dan *Clean Text* (teks yang telah dibersihkan).

3. Pra-pemrosesan Data

Pembersihan teks dilakukan dengan fungsi *clean_text* berbasis ekspresi reguler yang mengubah seluruh huruf menjadi huruf kecil (*case folding*), menghapus URL, *mention*, dan *hashtag*, menghapus semua karakter selain huruf Latin (a–z) dan spasi, merapikan spasi berlebih, serta membuang baris yang menjadi kosong. Selanjutnya, di RapidMiner dilakukan tokenisasi (memecah teks menjadi kata), penyaringan *stopword* menggunakan operator *Filter Stopwords*, dan *stemming* menggunakan algoritma *Snowball*. Tahapan pra-pemrosesan semacam ini juga digunakan pada klasifikasi sentimen ulasan aplikasi dan media sosial X (Agistina *et al.*, 2026; Fany *et al.*, 2026; Nida *et al.*, 2026). Contoh hasil pembersihan ditunjukkan pada Tabel 2.

Tabel 2. Contoh Teks Sebelum dan Sesudah Pembersihan

Sebelum	Sesudah
I think she grew a leg lol #sora #viral #wtf #grandma	*i think she grew a leg lol*
A man experiences a girlfriend robot. #ai #humanoid #robot #2025	*a man experiences a girlfriend robot*
Soo no 2nd appointment? #sora #ai #chiropractor #fyp #funnyvideos	*soo no nd appointment*
AI IS CRAZY	*ai is crazy*

4. Pembobotan Fitur dan Pelabelan

Teks diubah menjadi representasi numerik menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF) melalui operator *Process Documents from Data* dengan opsi *vector creation* TF-IDF. Pembobotan TF-IDF juga digunakan sebagai ekstraksi fitur pada klasifikasi teks ulasan aplikasi (Agistina *et al.*, 2026; Fany *et al.*, 2026). TF-IDF memberi bobot tinggi pada kata yang sering muncul dalam suatu dokumen tetapi jarang muncul di seluruh korpus, dengan bentuk umum:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right) \quad (1)$$

Dengan $\text{TF}(t, d)$ adalah frekuensi kata t pada dokumen d, N adalah jumlah dokumen, dan $\text{DF}(t)$ adalah jumlah dokumen yang memuat kata t. Pelabelan dilakukan pada operator *Generate Attributes* dengan aturan kata kunci: teks yang memuat istilah terkait AI, antara lain "*ai*", "*ai generated*", "*sora*", "*artificial intelligence*", dan "kecerdasan buatan", diberi label AI-*Generated*, sedangkan teks lainnya diberi label Non-AI.

5. Pembagian Data

Data dibagi menggunakan operator *Split Data* dengan rasio 70:30. Dari 1.008 data, diperoleh 705 data latih dan 303 data uji.

6. Pemodelan dengan *Logistic Regression*

Logistic Regression memodelkan hubungan antara variabel respons biner dan satu atau lebih fitur. Kombinasi linear dari fitur TF-IDF dihitung sebagai:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

Dengan z adalah kombinasi linear, β_0 adalah intersep, β_i adalah koefisien regresi, dan x_i adalah fitur ke-i (bobot TF-IDF dari teks *hashtag* dan komentar). Nilai z kemudian ditransformasikan dengan fungsi sigmoid menjadi probabilitas:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-z}} \quad (3)$$

Data diklasifikasikan sebagai kelas positif jika $P(y = 1 \mid x) \geq 0{,}5$ dan sebagai kelas negatif jika $P(y = 1 \mid x) < 0{,}5$. Koefisien diestimasi melalui optimasi iteratif untuk meminimalkan kesalahan prediksi. Alur pemodelan di RapidMiner terdiri atas operator *Read Excel*, *Set Role*, *Select Attributes*, *Generate Attributes*, *Nominal to Text*, *Process Documents from Data*, *Split Data*, *Logistic Regression*, *Apply Model*, dan *Performance* (Gambar 2).

7. Evaluasi Model

Kinerja model pada data uji diukur menggunakan *confusion matrix* yang memuat *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Evaluasi menggunakan *confusion matrix* juga

diterapkan pada penelitian klasifikasi teks sejenis (Agistina *et al.*, 2026; Fany *et al.*, 2026). Metrik evaluasi dihitung untuk masing-masing kelas sebagai berikut:

$$\text{Akurasi} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{4}$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{6}$$

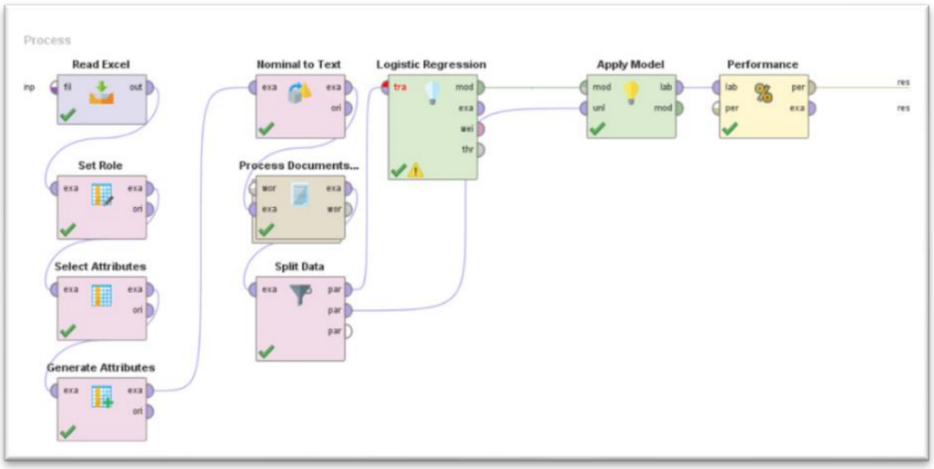
$$\text{F1} = \frac{2\times\text{Presisi}\times\text{Recall}}{\text{Presisi}+\text{Recall}} \tag{7}$$

Kelas AI-*Generated* dan kelas Non-AI masing-masing dievaluasi sebagai kelas positif, sehingga TN pada satu kelas sama dengan TP pada kelas lainnya.

3. Hasil dan Pembahasan

1. Hasil Implementasi Model

Gambar 2 menunjukkan alur pemodelan yang dibangun di RapidMiner. Alur dimulai dengan membaca *dataset* melalui *Read Excel*, mengubah peran atribut melalui *Set Role*, membuat label melalui *Generate Attributes*, mengubah tipe data dengan *Nominal to Text*, memproses dokumen menjadi vektor TF-IDF, membagi data latih dan data uji, melatih *Logistic Regression*, menerapkan model pada data uji, dan mengevaluasi kinerja dengan operator *Performance*.



Gambar 2. Proses Pemodelan dan Pengujian di RapidMiner

Dataset akhir berjumlah 1.008 data (705 data latih dan 303 data uji). Berdasarkan *confusion matrix*, data uji terdiri atas 195 data Non-AI (64,36%) dan 108 data AI-*Generated* (35,64%), sehingga distribusi kelas tidak seimbang.

2. Hasil Eksperimen

Hasil pengujian pada 303 data uji disajikan pada Tabel 3. Model memprediksi dengan benar 169 data Non-AI dan 50 data AI-*Generated*, sehingga akurasi keseluruhan adalah $\frac{169+50}{303} = 72{,}28\%$. Terdapat 84 kesalahan klasifikasi, terdiri atas 58 *false negative* (data AI-*Generated* yang diprediksi Non-AI) dan 26 *false positive* (data Non-AI yang diprediksi AI-*Generated*).

Tabel 3. *Confusion Matrix* pada Data Uji

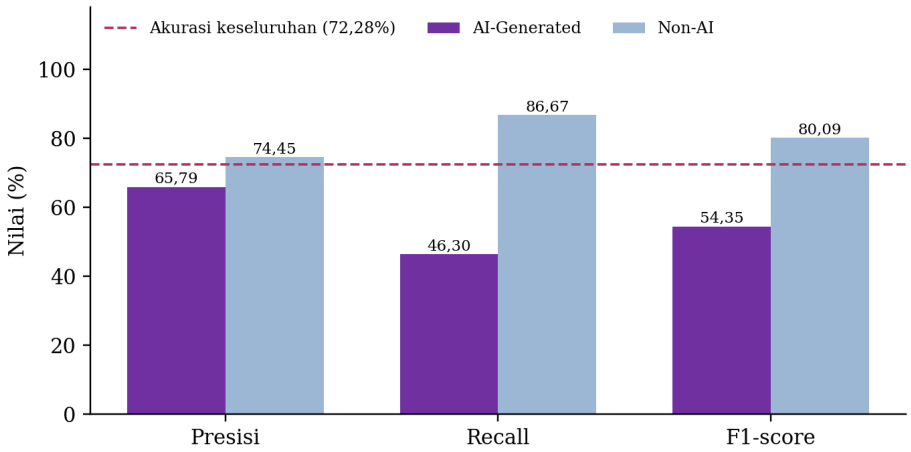
	Aktual Non-AI	**Aktual AI-*Generated***	**Presisi kelas**
Prediksi Non-AI	169	58	74,45%
Prediksi AI-*Generated*	26	50	65,79%
Recall kelas	86,67%	46,30%	—

Tabel 4. Metrik Evaluasi per Kelas

Kelas	**Presisi**	***Recall***	***F1-score***
AI-*Generated*	65,79%	46,30%	54,35%
Non-AI	74,45%	86,67%	80,09%
Rata-rata makro	70,12%	66,49%	67,22%

3. Visualisasi Hasil

Perbandingan metrik antarkelas ditampilkan pada Gambar 3. Model lebih sensitif terhadap kelas Non-AI, sedangkan *recall* dan *F1-score* kelas AI-*Generated* tertinggal jauh di bawah akurasi keseluruhan.



Gambar 3. Perbandingan Presisi, *Recall*, dan *F1-score* per Kelas

Pembahasan

Akurasi 72,28% lebih tinggi daripada akurasi pembanding sederhana yang selalu memprediksi kelas mayoritas (Non-AI), yaitu 64,36%, dengan selisih sekitar 7,9 poin persentase. Hal ini menunjukkan bahwa model menangkap sinyal leksikal yang membedakan kedua kelas, tetapi peningkatannya masih moderat. Pola kesalahan memperlihatkan bahwa model cenderung konservatif: 227 dari 303 data uji diprediksi Non-AI, sedangkan kelas tersebut hanya mencakup 195 data aktual.

Model bekerja baik pada kelas Non-AI dengan *recall* 86,67% (169 dari 195 data) dan presisi 74,45%. Pada kelas AI-*Generated*, model hanya mengenali 50 dari 108 data (*recall* 46,30%), sedangkan 58 data lainnya lolos sebagai *false negative*. Kesalahan jenis ini mencakup sekitar 69% dari seluruh kesalahan (58 dari 84). Ketika model memprediksi suatu data sebagai AI-*Generated*, prediksi tersebut benar pada 65,79% kasus (50 dari 76). Nilai *F1-score* kelas AI-*Generated* (54,35%) yang jauh di bawah kelas Non-AI (80,09%) menegaskan bahwa kelemahan utama model terletak pada sensitivitas terhadap kelas minoritas.

Akurasi pada penelitian ini lebih rendah daripada capaian Rendra *et al.* (2024) dan Rizal (2024) yang berada di atas 80%, serta capaian Alexander *et al.* (2023) dan Triyono *et al.* (2023) yang berada pada rentang 75–85%. Namun, perbandingan tersebut perlu dilakukan dengan hati-hati karena penelitian terdahulu menangani klasifikasi sentimen atau berita hoaks, sedangkan penelitian ini mengklasifikasikan asal konten (AI-*Generated* atau Non-AI) dari teks yang lebih pendek dan informal. *F1-score* kelas Non-AI (0,80) sebanding dengan *F1-score* 0,79 yang dilaporkan Cahyani *et al.* (2025) pada data tidak seimbang, tetapi *F1-score* kelas AI-*Generated* (0,54) masih jauh lebih rendah. Sebagai pembanding tambahan, akurasi 72,28% pada penelitian ini lebih rendah daripada *Naïve Bayes* pada ulasan App Store (79,78%; Agistina *et al.*, 2026) serta *Naïve Bayes* dan SVM pada teks media sosial X (masing-masing 80,5% dan 94,2%; Nida *et al.*, 2026), tetapi lebih tinggi daripada K-NN pada ulasan aplikasi DANA dengan tiga kelas sentimen (53,10%; Fany *et al.*, 2026). Rentang yang lebar ini menunjukkan bahwa akurasi sangat dipengaruhi oleh jumlah kelas, ukuran data, dan karakteristik teks, serta bahwa SVM layak diuji pada penelitian lanjutan. Temuan April *et al.* (2025) bahwa SVM sesuai untuk teks komentar TikTok yang pendek, serta temuan Guo *et al.* (2023) bahwa detektor berbasis *deep learning* lebih tangguh, mengindikasikan bahwa perbandingan dengan algoritma yang lebih kuat layak diuji, tetapi hal tersebut berada di luar batasan penelitian ini.

Beberapa faktor berpotensi memengaruhi hasil. Pertama, distribusi kelas yang tidak seimbang (sekitar 64% Non-AI dan 36% AI-*Generated* pada data uji) mendorong model condong ke kelas mayoritas. Kedua, ukuran *dataset* yang relatif kecil (1.008 data) dan teks yang sangat pendek membatasi informasi yang tersedia bagi TF-IDF, yang hanya menangkap frekuensi kata tanpa konteks. Ketiga, tahap pembersihan menghapus seluruh *hashtag* dari teks, padahal *hashtag* merupakan salah satu fitur yang dikaji; dampaknya terhadap sensitivitas kelas AI-*Generated* perlu diuji pada penelitian lanjutan. Keempat, alur pada Gambar 2 tidak memuat operator optimasi parameter, sehingga model dilatih tanpa penyetelan *hyperparameter*. Kelima, pelabelan berbasis kata kunci menentukan definisi kelas dan menjadi faktor kunci kualitas data.

Kelebihan penelitian ini adalah penggunaan metode yang ringan, dapat direplikasi tanpa pemrograman, dan koefisiennya dapat diinterpretasikan. Keterbatasannya meliputi: (a) label dibentuk dari aturan kata kunci pada teks yang sama dengan sumber fitur, sehingga terdapat risiko sirkularitas antara label dan fitur, dan penelitian ini tidak melaporkan validasi label oleh penilai manusia; (b) evaluasi menggunakan satu kali pembagian data tanpa validasi silang; (c) ketidakseimbangan kelas belum ditangani; (d) data hanya berasal dari TikTok; dan (e) tidak ada perbandingan dengan algoritma lain sesuai batasan masalah.

Secara praktis, model ini dapat menjadi *baseline* penyaringan awal konten yang berpotensi AI-*Generated* berdasarkan teks, dengan catatan bahwa konten yang lolos (*false negative*) masih cukup banyak sehingga hasilnya perlu dipadukan dengan verifikasi lain. Secara akademik, hasil ini menunjukkan bahwa fitur leksikal TF-IDF saja belum cukup sensitif dan perlu diperkaya dengan penanganan ketidakseimbangan kelas, validasi label yang independen, serta fitur lain.

4. Kesimpulan

Penelitian ini menunjukkan bahwa pengenalan konten AI-*Generated* dan Non-AI pada TikTok dapat dilakukan melalui pra-pemrosesan teks *hashtag* dan komentar (*case folding*, penghapusan URL, *mention*, *hashtag*, dan karakter non-alfabet, tokenisasi, penyaringan *stopword*, dan *stemming*). Metode *Logistic Regression* dengan pembobotan TF-IDF dapat digunakan untuk mengklasifikasikan kedua jenis konten, dengan akurasi keseluruhan 72,28% pada 303 data uji. Model lebih efektif mengenali kelas Non-AI (presisi 74,45%, *recall* 86,67%, *F1-score* 80,09%) dibandingkan kelas AI-*Generated* (presisi 65,79%, *recall* 46,30%, *F1-score* 54,35%), dan rendahnya *recall* tersebut disebabkan oleh 58 kasus *false negative*. Keterbatasan utama penelitian ini adalah ketidakseimbangan kelas, ukuran data yang kecil, pelabelan berbasis kata kunci, dan belum adanya penyetelan parameter. Penelitian selanjutnya disarankan menangani ketidakseimbangan kelas, memvalidasi label secara independen, membandingkan *Logistic Regression* dengan algoritma lain seperti SVM atau pendekatan *deep*

learning, memperluas dan memperbarui *dataset* secara berkala untuk menangkap perubahan kata dan *hashtag* di TikTok, serta mengintegrasikan fitur teks dengan fitur video.

5. Pernyataan Etik

Penelitian ini tidak melibatkan subjek manusia maupun hewan sehingga tidak memerlukan persetujuan etik (*ethical clearance*). Data berupa teks *hashtag* dan komentar yang tersedia secara publik di TikTok; seluruh *mention* dihapus pada tahap pembersihan dan tidak ada identitas pengguna yang dilaporkan.

6. Kontribusi Penulis

Khalifah Safa Rizky: konseptualisasi, metodologi, perangkat lunak, kurasi data, analisis formal, investigasi, visualisasi, penulisan naskah asli. Irwansyah: supervisi, validasi, peninjauan, dan penyuntingan naskah. Tupan Tri M: peninjauan dan penyuntingan naskah.

7. Pendanaan

Penelitian ini tidak menerima pendanaan khusus dari lembaga pemberi dana mana pun, baik di sektor publik, komersial, maupun nirlaba.

8. Pernyataan Konflik Kepentingan

Peneliti menyatakan tidak terdapat konflik kepentingan terkait penelitian, kepenulisan, dan/atau publikasi artikel ini.

9. Pernyataan Penggunaan AI Generatif

Penulis menggunakan AI untuk membantu menyusun penyuntingan bahasa. Data, analisis, hasil, dan kesimpulan berasal dari penelitian penulis; penulis telah meninjau dan bertanggung jawab penuh atas seluruh isi naskah.

Daftar Pustaka

Abidin, M. I., Nurtanio, I., & Achmad, A. (2022). Deepfake detection in videos using long short-term memory and CNN ResNext. *Jurnal Ilmiah ILKOM, 14*(3), 178–185. https://doi.org/10.33096/ilkom.v14i3.1254.178-185

Agistina, S. D., Irwansyah, I., Budiyantara, A., & Irwiensyah, F. (2026). Klasifikasi metode Naïve Bayes pada ulasan pengguna aplikasi Dazzcam untuk pengeditan foto vintage di App Store. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(2). https://doi.org/10.66217/digintel-ai.v1i2.9

Alexander, S. G., Ananto, A. T., Adhitya, I. P., Mangku, P., & Liano, B. (2023). Analisis sentimen opini masyarakat Indonesia terhadap konten deepfake tokoh publik. *KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer), 5*(2), 95–102. https://doi.org/10.54367/kakifikom.v5i2.3496

Alonso-López, N., Sidorenko-Bautista, P., & Giacomelli, F. (2021). Beyond challenges and viral dance moves: TikTok as a vehicle for disinformation and fact-checking in Spain, Portugal, Brazil, and USA. *Anàlisi: Quaderns de Comunicació i Cultura, 64*, 65–84. https://doi.org/10.5565/rev/analisi.3411

April, V. N., Manalu, P. D., Simanjuntak, M., & Umri, C. (2025). Implementasi algoritma klasifikasi untuk analisis sentimen media sosial TikTok tahun 2025. *JUTITI, 5*(2). https://doi.org/10.55606/jutiti.v5i2.5644

Cahyani, O. N., *et al.* (2025). Performa logistic regression dan naive Bayes dalam klasifikasi berita hoax di Indonesia. *Edumatic: Jurnal Pendidikan Informatika, 9*(1), 60–68. https://doi.org/10.29408/edumatic.v9i1.28987

Daulay, S. H., Saragih, A. S., & Bidadari, L. (2025). Linguistic strategies in the dissemination of hoaxes on TikTok: The role of hashtags in shaping public … *Journal of Applied Linguistic and Studies of Cultural, 2*(2), 9–19. https://doi.org/10.65787/jalc.v3i2.556

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. https://doi.org/10.48550/arXiv.1810.04805

Fadhilah, R. R. N. (2023). Dampak penyalahgunaan deepfake dalam memanipulasi informasi. *Jurnal Kawistara*. https://doi.org/10.22146/kawistara.98339

Fany, K. D., Irwansyah, I., Shidqon, M., & Pratiwi, N. (2026). Perbandingan kinerja algoritma K-Nearest Neighbor dan Decision Tree dalam analisis sentimen ulasan aplikasi DANA pada Google Play Store. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(2). https://doi.org/10.66217/digintel-ai.v1i2.11

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv*. https://doi.org/10.48550/arXiv.2301.07597

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *Proceedings of the 40th International Conference on Machine Learning, 202*, 24950–24962. https://doi.org/10.48550/arXiv.2301.11305

Nida, A. S., Irwansyah, I., Wiranata, A. D., & Halim, Z. (2026). Perbandingan Naive Bayes Classifier dan SVM untuk analisis sentimen desain seragam atlet Indonesia pada media sosial X di Olimpiade Paris 2024. *DIGINTEL-AI: DIGital INnovation and inTELligence – AI, 1*(2). https://doi.org/10.66217/digintel-ai.v1i2.13

Nirmala, N. M. G. S., & Sanjaya E. R., N. A. (2024). Analisis sentimen dengan logistic regression untuk deteksi kata pada Livin' by Mandiri. *Jurnal Nasional Teknologi Informasi dan Aplikasinya, 2*(4), 869–878. https://doi.org/10.24843/JNATIA.2024.v02.i04.p25

Putri, A. N., Prasetyo, R. A., & Crisnawati, C. I. G. (2026). Analisis sentimen terhadap AI dalam pendidikan di media sosial menggunakan metode kuantitatif logistic regression. *Jurnal Elektro dan Informatika Swadharma (JEIS), 6*(1), 69–77. https://doi.org/10.56486/jeis.vol6no2.1129

Rahayu, N. A., & Sudjradjat, R. H. (2025). The influence of non-AI and AI-generated TikTok true story content on Generation Z's perceptions (A study on elaboration likelihood theory). *INJECT: Interdisciplinary Journal of Communication, 10*(1), 821–846. https://doi.org/10.18326/inject.v10i1.4504

Rendra, S., *et al.* (2024). Optimasi model machine learning untuk analisis sentimen data komentar TikTok tentang kemajuan Ibu Kota Nusantara. *Journal of Informatics and Vocational Education, 5*(4), 200–215. https://doi.org/10.46336/ijmsc.v3i3.232

Rizal, F. (2024). Analisis sentimen masyarakat Indonesia terhadap aplikasi TikTok menggunakan logistic regression. *Jurnal Akiratech, 2*(1), 20–30. https://doi.org/10.63935/akiratech.v1i2.33

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. https://doi.org/10.48550/arXiv.2303.11156

Setyawan, A., Fitriani, A., & Rilvani, E. (2025). Klasifikasi kemiskinan di Indonesia dengan decision tree menggunakan RapidMiner. *Jurnal Media Akademik (JMA), 3*(7). https://doi.org/10.62281/v3i7.2646

Triyono, L., Gernowo, R., Rahaman, M., & Raharjo, T. (2023). Indonesian fake news detection using various machine learning technique. *International Journal on Informatics Visualization, 7*(3), 726–732. https://doi.org/10.30630/joiv.7.3.1243

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TuringBench: A benchmark environment for Turing test in the age of neural text generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2001–2016. https://doi.org/10.18653/v1/2021.findings-emnlp.172

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics, 51*(1). https://doi.org/10.1162/coli_a_00549