Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

DETEKSI TEKS BERITA GENERASI AI BERDASARKAN FITUR STILOMETRI MENGGUNAKAN TABPFN DAN SHAP

Syamsul Alam[1]*, Muhammad Faisal [2], Rizki Yusliana Bakti [3]
[1,2,3]Universitas Muhammadiyah Makassar, Indonesia
[1]105841117922@student.unismuh.ac.id, [2]muhfaisal@unismuh.ac.id ,
[3]rizkiyusliana@unismuh.ac.id
*Coresponding Author

Info Artikel

Pengajuan : 22-08-2026
Peninjauan : 23-08-2026
Revisi : 08-09-2026
Diterima : 23-08-2026
Diterbitkan : 15-09-2026

Kata Kunci:
Deteksi Berita, Explainable AI, SHAP, Stilometri, TabPFN, Teks Sintesis AI.

Abstrak

Kemajuan pesat Large Language Models (LLM) seperti Llama 3.1 memicu eskalasi disinformasi digital akibat kemampuannya memproduksi teks berita sintetis yang menyerupai gaya penulisan jurnalis manusia. Penelitian ini mengusulkan kerangka deteksi forensik digital teks berbasis fitur stilometri tingkat permukaan yang dipadukan dengan algoritma klasifikasi Prior-Data Fitted Networks (TabPFN) serta transparansi model berbasis Shapley Additive Explanations (SHAP). Eksperimen dilakukan menggunakan korpus seimbang 500 dokumen berita berbahasa Indonesia (250 artikel orisinal Kompas.com dan 250 teks sintesis Llama 3.1) yang dipartisi secara terstratifikasi dengan rasio 80:20, yaitu 400 data latih dan 100 data uji. Enam atribut stilometri diekstraksi tanpa melalui tahap pembuangan kata henti maupun perataan morfologi guna mempertahankan keaslian sidik jari gaya kepenulisan. Pemanfaatan TabPFN memungkinkan inferensi data tabular dieksekusi secara instan melalui mekanisme in-context learning tanpa membutuhkan penalaan hiperparameter yang intensif. Hasil evaluasi empiris membuktikan keandalan tinggi dan konsistensi klasifikasi model dalam membedakan naskah berita manusia dari teks buatan mesin secara objektif dan independen terhadap topik wacana. Analisis keterbukaan model menggunakan SHAP mengungkap bahwa Type-Token Ratio (TTR) merupakan determinan stilometri paling krusial, di mana tingginya diversitas kosakata jurnalis manusia menjadi faktor pembeda esensial terhadap kecenderungan repetisi leksikal mesin. Secara keilmuan, integrasi ini membuktikan efektivitas pendekatan stilometri terukur sebagai instrumen forensik digital teks yang tangguh, efisien secara komputasi, dan akuntabel dalam mendukung verifikasi keaslian naskah pada ekosistem jurnalisme media daring.

How to Cite : Alam, S., Faisal, M., & Bakti, R. Y. (2026). Deteksi teks berita generasi AI berdasarkan fitur stilometri menggunakan TabPFN dan SHAP. *Journal of Computer Science and Information Technology (JCSIT), 3*(4), 399-410.

DOI : 10.70248/jcsit.v3i4.4557

PENDAHULUAN

Kemampuan representasi teks oleh *Large Language Models* (LLM) mutakhir, seperti OpenAI GPT-3.5 dan Llama 3.1, telah merevolusi ekosistem generasi konten otomatis dengan menghasilkan narasi artikel yang memiliki kelancaran sintaksis dan keselarasan gramatikal menyerupai karya jurnalis profesiona (Mitchell et al., 2023),(Grattafiori et al., 2024). Kemajuan ini menghadirkan tantangan kritis terhadap integritas ruang informasi digital, terutama terkait potensi fabrikasi fakta otomatis, disinformasi berskala masif, serta sulitnya memverifikasi keaslian naskah berita yang diterbitkan (Putu & Antari, 2025),(Narendra et al., 2024). Identifikasi teks hasil generasi AI secara

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

kasatmata sangat sulit dilakukan oleh manusia karena model generatif saat ini telah mampu mematuhi tata bahasa formal dan meminimalkan kesalahan tipografi (Of et al., 2025). Oleh karena itu, pengembangan metode komputasi cerdas yang mampu mendeteksi jejak digital kepenulisan teks secara otomatis, akurat, dan objektif menjadi kebutuhan yang sangat krusial.

Penelitian terdahulu mengenai deteksi teks kecerdasan buatan umumnya bertumpu pada pemodelan semantik mendalam melalui representasi vektor kata (*embeddings*) dan penyelarasan (*fine-tuning*) model Transformer berskala besar seperti BERT atau RoBERTa (Grattafiori et al., 2024), (Sa & Gerhana, 2025). Meskipun pendekatan berbasis semantik ini mencatatkan performa klasifikasi yang tinggi pada data uji yang homogen, model Transformer mendalam memiliki kelemahan komputasi dan metodologis yang substansial (Hollmann, Müller, Purucker, Krishnakumar, et al., 2025). Penyelarasan arsitektur Transformer dengan puluhan hingga ratusan juta parameter menuntut alokasi sumber daya komputasi grafis (GPU) yang masif, jejak memori (*memory footprint*) yang besar, serta biaya komputasi yang sangat mahal (Hollmann, Müller, Purucker, Krishnakumar, et al., 2025), . Selain itu, proses inferensi representasi semantik memerlukan waktu latensi yang relatif tinggi sehingga tidak efisien jika diimplementasikan pada perangkat berspesifikasi standar atau sistem verifikasi berita waktu-nyata (*real-time streaming*) (Hollmann, Müller, Purucker, Krishnakumar, et al., 2025). Dari sisi representasi data, model semantik sangat rentan terhadap bias dependensi topik (*topic-dependence bias*) dan masalah lewat-adaptasi (*overfitting*) terhadap entitas faktual. Representasi kontekstual Transformer cenderung menangkap korelasi kata benda, nama tokoh, atau tema peristiwa spesifik yang dominan pada data latih alih-alih mengisolasi pola gaya kepenulisan murni. Keterikatan semantik ini menyebabkan model mengalami penurunan akurasi generalisasi secara drastis (*generalization drop*) saat diuji pada artikel berita dengan topik peristiwa terkini di luar distribusi data latih (*out-of-distribution*) (Yatheendra & Arabagatte, 2025), (Emekci & Özkan, 2025).

Sebagai alternatif yang jauh lebih efisien secara komputasi dan sepenuhnya independen terhadap topik wacana, pendekatan analisis stilometri tingkat permukaan (*surface-level stylometry*) menawarkan solusi yang tangguh [9], [10]. Ekstraksi fitur stilometri beroperasi dengan biaya komputasi yang sangat rendah karena hanya membutuhkan operasi pemindaian teks berbasis CPU dalam hitungan milidetik tanpa memerlukan infrastruktur akselerator grafis yang kompleks [10]. Pendekatan ini mengukur karakteristik kuantitatif kebiasaan linguistik alami teks, seperti panjang kalimat, kompleksitas morfologi kata, kepadatan leksikal, dan intensitas penggunaan tanda baca (Tran et al., 2024),(Bagies et al., 2025). Variabel-variabel stilometri ini merefleksikan sidik jari gaya penulisan (*stylistic fingerprint*) yang membedakan cara jurnalis manusia menyusun wacana berita dari pola probabilistik yang diproduksi oleh model bahasa generatif(Bagies et al., 2025),(Tanaja et al., 2025). Karakteristik ini sangat relevan pada korpus berita berbahasa Indonesia, di mana kaidah penulisan jurnalistik lokal menuntut struktur kalimat yang padat, lugas, dan dinamis, sementara teks sintesis model generatif kerap memperlihatkan repetisi kata tugas formal, struktur afiksasi bertingkat yang kaku, serta konstruksi kalimat yang cenderung seragam (Tanaja et al., 2025).

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

Dalam mengolah data berstruktur tabular dari hasil ekstraksi fitur stilometri, sebagian besar penelitian sebelumnya menerapkan algoritma pembelajaran mesin tradisional seperti Random Forest, Support Vector Machines (SVM), atau XGBoost (Bagies et al., 2025),(Tanaja et al., 2025). Akan tetapi, algoritma-algoritma tersebut membutuhkan proses penalaan hiperparameter (*hyperparameter tuning*) yang intensif agar performa klasifikasi tidak rentan terhadap kondisi lewat-latih (*overfitting*) pada ukuran dataset moderat (Tanaja et al., 2025). Untuk mengatasi batasan komputasi tersebut, penelitian ini memanfaatkan algoritma *Prior-Data Fitted Networks* (TabPFN), yaitu model berbasis Transformer yang telah dipra-latih menggunakan jutaan tugas sintetis sehingga mampu melakukan inferensi klasifikasi data tabular secara instan melalui mekanisme *in-context learning* tanpa membutuhkan penalaan hiperparameter yang rumit (Hollmann, Müller, Purucker, & Krishnakumar, 2025).

Selain keandalan metrik klasifikasi, tantangan fundamental lainnya dalam penerapan kecerdasan buatan pada ranah forensik teks digital adalah masalah ketertutupan model (*black-box problem*) (Patel et al., 2026),(Hamilton & Papadopoulos, 2023). Keputusan klasifikasi yang dihasilkan oleh algoritma sering kali tidak disertai dengan penalaran logis yang dapat dipertanggungjawabkan kepada pengguna, sehingga membatasi penerapannya dalam proses verifikasi informasi publik (Patel et al., 2026). Kebaruan dari penelitian ini terletak pada integrasi antara ekstraksi enam fitur stilometri tingkat permukaan, klasifikasi instan TabPFN, dan pembongkaran transparansi keputusan model menggunakan metode *Shapley Additive Explanations* (SHAP) (Hamilton & Papadopoulos, 2023),(Winarjo & Sari, 2026). Pendekatan ini tidak hanya menguji performa klasifikasi secara kuantitatif, tetapi juga membuktikan kontribusi matematis dari setiap variabel stilometri secara global maupun lokal pada dokumen berita yang diuji (Winarjo & Sari, 2026), (Ghosh & Khandoker, 2024).

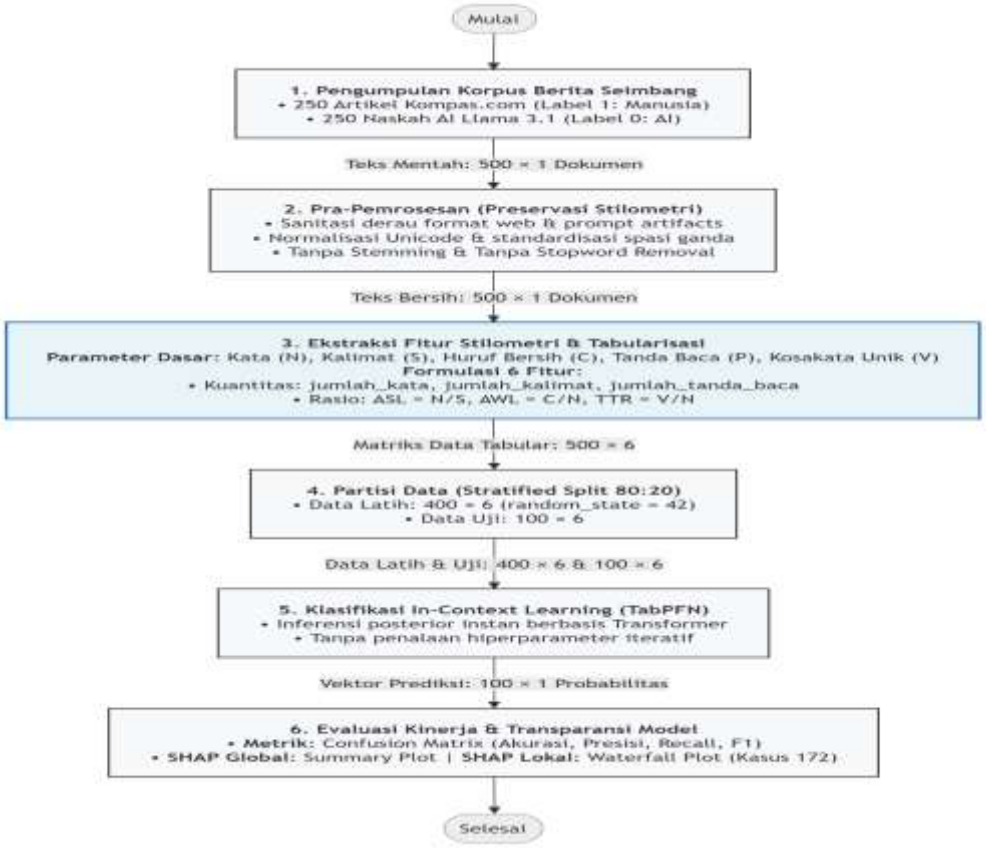
Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk merancang alur klasifikasi teks berita sintesis AI (Llama 3.1) dan berita asli manusia (*Kompas.com*) menggunakan perpaduan fitur stilometri dan model TabPFN, mengevaluasi kinerja klasifikasi berdasarkan metrik evaluasi standar, serta membedah pola atribusi keputusan model melalui interpretasi SHAP. Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan metode deteksi teks AI pada korpus berbahasa Indonesia, memberikan pemahaman mendalam mengenai pola linguistik pembeda teks, serta menjadi instrumen forensik digital yang akurat dan akuntabel dalam menjaga integritas dan keaslian informasi di era kecerdasan buatan.

METODE PENELITIAN

Penelitian ini dirancang menggunakan pendekatan eksperimental kuantitatif berbasis pembelajaran mesin untuk data tabular guna mengidentifikasi keaslian naskah berita berbahasa Indonesia secara objektif. Kerangka kerja perancangan sistem disusun secara terstruktur dan terbagi secara linier ke dalam enam tahapan komputasi utama. Rangkaian tahapan tersebut mencakup: (1) pengumpulan korpus berita seimbang yang menghimpun artikel orisinal jurnalis manusia dan teks berita sintetis buatan model bahasa besar, (2) pra-pemrosesan data berpedoman pada preservasi stilometri guna mempertahankan karakteristik

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

leksikal alami tanpa pembuangan kata henti maupun perataan morfologi, (3) pemrosesan atribut linguistik dasar serta ekstraksi enam fitur stilometri tingkat permukaan untuk mengonversi korpus teks tidak terstruktur menjadi matriks data tabular terstandarisasi, (4) partisi data secara terstratifikasi guna menjamin keseimbangan distribusi kelas pada subset latih dan uji, (5) klasifikasi probabilitas posterior instan menggunakan algoritma *Prior-Data Fitted Networks* (TabPFN) berbasis *in-context learning* tanpa membutuhkan penalaan hiperparameter yang rumit, serta (6) evaluasi komprehensif metrik performa klasifikasi yang diintegrasikan dengan pembongkaran transparansi keputusan model secara global maupun lokal menggunakan *Shapley Additive Explanations* (SHAP). Keseluruhan keterhubungan alur kerja sistem beserta dinamika transformasi dimensi matriks data pada setiap tahapannya diilustrasikan secara rinci.



Gambar 1. Alur Perancangan Sistem

Objek data dalam penelitian ini adalah korpus teks berita berbahasa Indonesia berjumlah 500 dokumen dengan komposisi seimbang (*balanced dataset*). Data kelas teks asli label 1 terdiri dari 250 artikel berita terverifikasi yang dihimpun dari portal media daring *Kompas.com*. Sementara itu, data kelas teks kecerdasan buatan label 0 terdiri dari 250 dokumen artikel berita yang disintesis menggunakan model *Large Language Model* Llama 3.1 dengan memberikan topik dan fakta kunci yang setara dengan korpus berita manusia.

Pemilihan artikel berita dikontrol secara ketat melalui kriteria inklusi dan eksklusi. Kriteria inklusi mencakup artikel berita lugas (*hard news*) berbahasa Indonesia dari kanal nasional *Kompas.com* yang diterbitkan dalam rentang waktu

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

Januari hingga Desember 2024, mencakup topik politik, ekonomi, sosial, dan teknologi dengan panjang berkisar antara 150 hingga 800 kata. Kriteria eksklusi mengeluarkan artikel opini, tajuk rencana (*editorial*), siaran pers promosi komersial, hasil terjemahan mesin, serta artikel dengan panjang di bawah 150 kata guna mempertahankan homogenitas struktur wacana faktual. Topik dan fakta kunci dari naskah jurnalis tersebut selanjutnya digunakan sebagai *prompt* bagi model Llama 3.1 untuk menghasilkan 250 naskah pasangan dengan tema setara.

Tahap pra-pemrosesan data dilakukan dengan berpedoman ketat pada prinsip preservasi stilometri. Teks tidak melalui proses *stemming*, penghapusan kata henti (*stopword removal*), maupun pembuangan tanda baca, karena struktur-struktur tersebut merupakan representasi utama dari ciri gaya kepenulisan. Pembersihan data difokuskan pada sanitasi teknis yang mencakup normalisasi karakter *Unicode*, standardisasi spasi ganda, penghapusan sisa tautan artikel rujukan dari media web, serta pembersihan teks pembuka atau penutup bawaan dari keluaran *prompt* model Llama 3.1. Tahap ini juga menjalankan validasi data untuk memastikan tidak adanya dokumen kosong atau duplikasi baris teks.

Ekstraksi Fitur Stilometri

Korpus teks yang telah bersih ditransformasikan ke dalam matriks tabular berdimensi 500×6 melalui ekstraksi enam fitur stilometri tingkat permukaan. Variabel yang diekstraksi mencakup jumlah kata, jumlah kalimat, jumlah tanda baca, *Average Sentence Length* (ASL), *Average Word Length* (AWL), dan *Type-Token Ratio* (TTR).

Tabel 1. Fitur Stilometrik yang Diekstraksi dan Hipotesis Pembeda

Kode Fitur	Nama Fitur	Definisi	Hipotesis Perbedaan AI vs Manusia
ASL	*Average Sentence Length*	Σ kata / Σ kalimat	Teks AI distribusi panjang kalimat lebih seragam; Teks Manusia lebih bervariasi.
AWL	*Average Word Length*	Σ karakter / Σ kata	Teks AI cenderung menggunakan diksi formal dan lebih panjang secara konsisten.
TTR	*Type-Token Ratio*	Jumlah kata unik / Total kata	Teks AI memiliki TTR lebih rendah (pengulangan kosakata lebih tinggi).
JKT	Jumlah Kata	Total token dalam dokumen	Teks AI cenderung lebih konsisten dalam panjang artikel.
JKL	Jumlah Kalimat	Total kalimat (pemisah: . ! ?)	Proksi kepadatan informasi per artikel.
JTB	Jumlah Tanda Baca	Frekuensi . , ! ? : ; dalam dokumen	Teks AI penggunaan tanda baca lebih sistematis dan teratur.

Formulasi matematis untuk menghitung ketiga rasio stilometri tersebut dinyatakan dalam persamaan berikut:

$$\text{ASL} = \frac{N}{S} \quad (1)$$

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

$$\text{AWL} = \frac{C}{N} \quad (2)$$

$$\text{TTR} = \frac{V}{N} \quad (3)$$

Di mana N merepresentasikan total kata, S adalah total kalimat, C menyatakan akumulasi panjang huruf alfabet bersih tanpa spasi dan simbol, serta V merupakan total kata unik (*vocabulary size*) pada dokumen.

Matriks data tabular dipersiapkan melalui penyesuaian tipe data numerik dan dipartisi menggunakan teknik *Stratified Train-Test Split* dengan proporsi 80% data latih (400 dokumen) dan 20% data uji (100 dokumen) serta penguncian acak *random state* 42. Proses klasifikasi dieksekusi menggunakan model TabPFN (*Prior-Data Fitted Networks*). Algoritma ini memanfaatkan arsitektur Transformer yang menerapkan inferensi *in-context learning* untuk langsung mengestimasi probabilitas posterior kelas biner data uji berdasarkan matriks data latih tanpa melibatkan tahapan penalaan hiperparameter iteratif.

Pengujian performa model klasifikasi diukur menggunakan *Confusion Matrix* data uji yang memetakan kuadran *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan matriks tersebut, efektivitas sistem dihitung secara kuantitatif melalui empat metrik evaluasi utama, yaitu akurasi, presisi, *recall*, dan *F1-score*.

Transparansi proses klasifikasi dibedah menggunakan metode *Shapley Additive Explanations* (SHAP). SHAP mengalokasikan nilai kontribusi marjinal (ϕ_i) untuk setiap fitur stilometri berdasarkan teori permainan kooperatif, sehingga nilai prediksi akhir model $f(x)$ merupakan jumlahan linier aditif dari nilai dasar ekspektasi $E[f(X)]$ dengan total kontribusi seluruh fitur:

$$f(x) = E[f(X)] + \sum_{i=1}^{M} \phi_i \quad (4)$$

Di mana M merupakan total fitur input ($M = 6$). Evaluasi transparansi dilakukan secara global melalui *SHAP Summary Plot* untuk mengidentifikasi urutan kepentingan fitur pada seluruh data uji, serta secara lokal melalui *SHAP Waterfall Plot* guna membuktikan pergeseran nilai probabilitas pada sampel uji studi kasus.

HASIL PENELITIAN DAN PEMBAHASAN

Pengumpulan dan Prapengolahan Data

Tahap pengumpulan dan pra-pemrosesan data berhasil menyusun korpus bersih berisi 500 dokumen berita berbahasa Indonesia dengan distribusi kelas seimbang, terdiri dari 250 artikel orisinal jurnalis *Kompas.com* dan 250 teks berita sintesis model Llama 3.1. Proses sanitasi teknis berhasil membersihkan seluruh derau format web dan artefak pembuka *prompt* LLM tanpa mengubah struktur tanda baca maupun variasi kata. Pemeriksaan integritas data memastikan seluruh

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

baris data lengkap tanpa adanya nilai kosong (*missing values*) ataupun duplikasi dokumen.

Hasil Ekstraksi Fitur Stilometri dan Validasi Perhitungan Manual

Validitas fungsi ekstraksi komputasi Python diverifikasi melalui pembuktian matematis manual pada sampel Dokumen 1 (Teks AI) yang memiliki parameter dasar: 327 kata (N), 19 kalimat (S), 2.067 karakter huruf bersih (C), 56 tanda baca (P), dan 162 kata unik (V). Perhitungan manual menghasilkan rasio $\text{ASL} = 327/19 = 17{,}211$, $\text{AWL} = 2.067/327 = 6{,}321$, dan $\text{TTR} = 162/327 = 0{,}495$. Nilai-nilai ini terkonfirmasi presisi dan identik hingga tiga tempat desimal dengan luaran skrip komputasi sistem, sehingga fungsi ekstraksi fitur stilometri dinyatakan valid secara matematis tanpa galat komputasi. Sebagian hasil ekstraksi fitur tersebut disajikan pada Tabel 2.

Hasil perhitungan manual membuktikan bahwa angka yang diproduksi secara komputasional oleh sistem pengekstraksi fitur bernilai sama persis dengan perhitungan matematis.

Setelah proses ekstraksi fitur selesai dilakukan, setiap dokumen direpresentasikan dalam bentuk vektor numerik yang memuat jumlah karakter, jumlah kata, jumlah kalimat, jumlah tanda baca, *Average Sentence Length* (ASL), *Average Word Length* (AWL), *Type-Token Ratio* (TTR).

Tabel 2. Hasil ektrafitur stilometri

NO	Jumlah Karakter	Jumlah Kata	Jumlah Kalimat	Jumlah Tanda Baca	ASL	AWL	TTR	Label
1	2444	327	19	56	17,211	6,321	0,495	0
2	3023	424	19	125	22,316	5,95	0,627	1
3	2867	360	16	52	22,5	6,828	0,419	0
4	2338	342	24	73	14,25	5,737	0,573	1
5	1727	243	15	65	16,2	5,971	0,638	1
…	…	…	…	…	…	…	…	…
496	1978	273	11	56	24,818	6,179	0,582	1
497	3038	427	22	70	19,409	5,979	0,389	0
498	2557	305	20	83	15,25	7,134	0,518	0
499	2954	389	15	67	25,933	6,465	0,411	0
500	2281	306	14	58	21,857	6,33	0,618	1

Label 1 menunjukkan dokumen yang termasuk dalam kelas teks yang ditulis oleh manusia, sedangkan label 0 menunjukkan dokumen yang termasuk dalam kelas teks hasil generasi AI

Data yang disajikan pada Tabel 2, merupakan representasi akhir yang siap digunakan sebagai masukan (*input*) bagi model klasifikasi TabPFN. Seluruh fitur telah direpresentasikan dalam bentuk numerik sehingga tidak memerlukan proses *encoding* tambahan.

Hasil Klasifikasi Menggunakan TabPFN dan Evaluasi Kinerja Model

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410

Pengujian klasifikasi terhadap 100 data uji yang terbagi secara terstratifikasi (50 data teks manusia dan 50 data teks AI) menghasilkan pemetaan matriks kontingensi.

Tabel 3. *Confusion Matrix* Model TabPFN

Aktual	**AI**	**Manusia**
AI	TP = 45	FN = 6
Manusia	FP = 5	TN = 44

Berdasarkan hasil pada Tabel 3, model TabPFN berhasil mengidentifikasi 44 dari 50 dokumen berita jurnalis manusia secara tepat (*True Positive*) dan mengenali 45 dari 50 teks berita sintesis AI secara benar (*True Negative*). Kesalahan prediksi tercatat sangat minimal, yakni 5 dokumen teks AI yang salah diklasifikasikan sebagai teks manusia (*False Positive*) serta 6 dokumen karya manusia yang terdeteksi sebagai teks AI (*False Negative*). Berdasarkan matriks kontingensi tersebut, kalkulasi empat metrik evaluasi kinerja model.

Tabel 4. Metrik Evaluasi *Model TabPFN*

Metrik Evaluasi	**Nilai**
Akurasi	89.0%
Presisi	89.7%
Recall	88.0%
F1-Score	88.8%

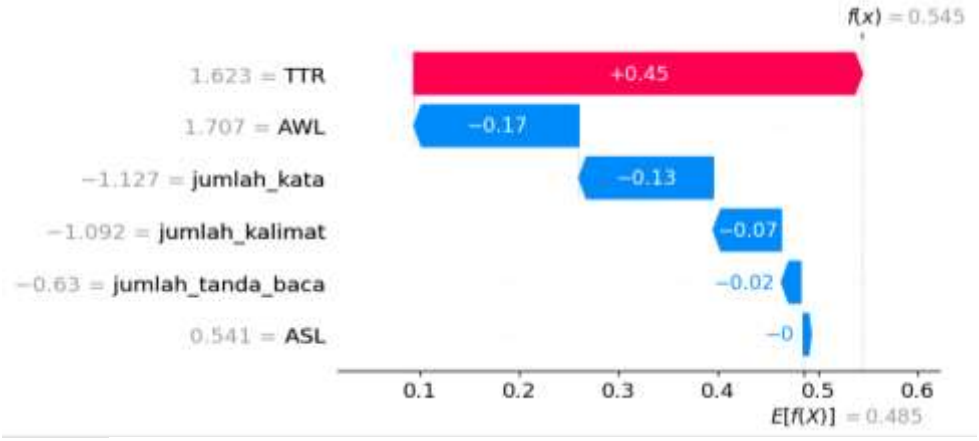
Hasil evaluasi menunjukkan bahwa kombinasi enam fitur stilometri dengan model TabPFN menghasilkan tingkat akurasi sebesar 89.0% dan nilai *F1-score* sebesar 88.8%. Capaian ini membuktikan bahwa pendekatan stilometri tingkat permukaan memiliki daya diskriminasi yang kuat untuk membedakan artikel berita manusia dari teks buatan mesin tanpa membutuhkan representasi semantik yang rumit.

Hasil Interpretasi Model Berbasis SHAP

Pengujian transparansi model secara global dilakukan menggunakan *SHAP Summary Plot* untuk menganalisis hierarki pengaruh seluruh fitur stilometri pada data uji. Hasil evaluasi global menunjukkan bahwa fitur *Type-Token Ratio* (TTR) menempati urutan kepentingan teratas dan menjadi variabel paling berpengaruh dalam pengambilan keputusan model. Nilai TTR yang tinggi secara konsisten memberikan nilai atribusi positif yang mendorong klasifikasi ke arah teks berita jurnalis manusia (Label 1), sedangkan nilai TTR yang rendah mendorong keputusan ke arah teks sintesis AI (Label 0). Fitur jumlah kata dan jumlah tanda baca menempati posisi pengaruh berikutnya, sementara ASL, jumlah kalimat, dan AWL memberikan kontribusi marjinal yang lebih moderat.

Untuk membuktikan transparansi keputusan pada level instansi individual, analisis lokal dievaluasi melalui *SHAP Waterfall Plot* pada sampel dokumen uji dengan nilai dasar ekspektasi data latih sebesar $E[f(X)] = 0.485$. Rincian kontribusi nilai marjinal setiap fitur.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 399-410



Gambar 2. SHAP Waterfall Plot untuk dokumen studi kasus

Kontribusi negatif yang cukup signifikan ditunjukkan oleh variabel Average Word Length ($\phi_{\mathrm{AWL}} = -0{,}17$, dengan nilai terstandarisasi 1,707 karakter). Secara linguistik komputasi bahasa Indonesia, fenomena ini memberikan wawasan mendasar mengenai karakteristik gaya berita lokal. Pada penulisan artikel *hard news* di portal media seperti *Kompas.com*, jurnalis berpegang teguh pada kaidah piramida terbalik yang menuntut paragraf pembuka (*lead*) ditulis secara ringkas, padat, dan komunikatif dengan klausa langsung. Sebaliknya, model generatif Llama 3.1 memperlihatkan kecenderungan bawaan (*generative bias*) untuk memilih kosakata formal turunan dengan struktur afiksasi bertingkat yang menghasilkan kata berkarakter panjang. Tingginya rata-rata panjang kata pada naskah individual ini dibaca oleh model sebagai karakteristik teks mesin sehingga menarik prediksi ke arah Label 0. Namun, dorongan positif masif dari variasi kosakata ($\phi_{\mathrm{TTR}} = +0{,}45$) berhasil mengompensasi tarikan negatif tersebut dan menetapkan dokumen pada kelas jurnalis manusia secara tepat.

KESIMPULAN

Penelitian ini membuktikan bahwa perpaduan analisis stilometri tingkat permukaan dan algoritma *Prior-Data Fitted Networks* (TabPFN) menyajikan kerangka deteksi teks berita sintesis AI yang andal, efisien secara komputasi, dan objektif pada korpus berbahasa Indonesia. Pemanfaatan mekanisme *in-context learning* pada TabPFN meniadakan kebutuhan optimasi gradien maupun penalaan hiperparameter manual yang rumit, sekaligus mengatasi keterbatasan komputasi model pemrosesan bahasa alami berbasis semantik mendalam yang menuntut sumber daya perangkat keras masif. Pendekatan stilometri tingkat permukaan ini terbukti tangguh karena beroperasi secara independen terhadap variasi topik wacana, sehingga terhindar dari bias entitas faktual yang kerap menurunkan performa generalisasi model generatif.

Pembongkaran transparansi keputusan model melalui SHAP secara konseptual menegaskan adanya polarisasi gaya kepenulisan yang tegas antara naskah orisinal jurnalis manusia dan teks hasil sintesis mesin. Variabel *Type-Token Ratio* (TTR) terbukti secara konsisten menjadi determinan gaya bahasa paling dominan dalam model. Tingginya diversitas kosakata jurnalis manusia merefleksikan fleksibilitas linguistik yang kontras dengan kecenderungan model

bahasa besar (seperti Llama 3.1) yang terperangkap dalam pengulangan token-token umum berprobabilitas tinggi. Di samping itu, kebiasaan generator AI yang memproduksi kata-kata formal panjang dengan struktur afiksasi bertingkat yang kaku (*Average Word Length* tinggi) secara konsisten diasosiasikan oleh model sebagai penanda teks sintetis. Integrasi ini membuktikan bahwa analisis gaya bahasa kuantitatif mampu menjadi instrumen forensik digital teks yang transparan, akuntabel, dan dapat dipertanggungjawabkan dalam menjaga integritas informasi publik.

Pengembangan penelitian selanjutnya direkomendasikan untuk menguji ketahanan generalisasi metode ini terhadap teks yang disintesis oleh model bahasa besar generasi termutakhir lainnya, seperti OpenAI GPT-4, Anthropic Claude 3.5 Sonnet, dan Google Gemini Pro. Selain itu, eksplorasi lanjutan dapat diperluas pada analisis variasi genre penulisan non-faktual (seperti kolom opini dan sastra) serta evaluasi performa pada korpus jurnalisme multibahasa.

Pernyataan mengenai Penggunaan Kecerdasan Buatan (AI).

AI digunakan dalam penyusunan naskah ini. Penulis menggunakan perangkat kecerdasan buatan semata-mata untuk membantu perbaikan bahasa, pemeriksaan tata bahasa, serta penyuntingan demi kejelasan dan keterbacaan. Penulis tetap bertanggung jawab penuh atas akurasi, orisinalitas, integritas, dan konten akademis naskah tersebut. Seluruh konten yang dibantu oleh AI telah ditinjau dan diverifikasi oleh penulis sebelum naskah diserahkan.

Konflik kepentingan

Penulis menyatakan tidak ada konflik kepentingan.

Kontribusi penulis

Konseptualisasi, S.A. dan M.F.; pengumpulan data, S.A.; validasi data, M.F.; analisis formal, S.A. dan R.Y.B.; penulisan-penyusunan draf awal, S.A.; penulisan-peninjauan dan penyuntingan, M.F. dan R.Y.B.; visualisasi, S.A. Seluruh penulis telah membaca dan menyetujui versi akhir naskah ini.

Ucapan terima kasih

Penelitian ini didukung oleh Program Studi Informatika, Fakultas Teknik, serta Lembaga Penelitian, Publikasi, dan Pengabdian Masyarakat (LP3M) Universitas Muhammadiyah Makassar. Penulis mengucapkan terima kasih atas dukungan fasilitas komputasi, akses literatur, dan bimbingan akademik yang diberikan untuk memfasilitasi penyelesaian penelitian ini.

DAFTAR PUSTAKA

Bagies, T., Alsuhaimi, R., Almasre, M., & Bafail, A. (2025). Predicting Suspicious Arabic X Accounts Using Stylometric Features. *IEEE Access*, *13*(September), 153464–153473. https://doi.org/10.1109/ACCESS.2025.3605126

Emekci, H., & Özkan, İ. (2025). Stylometric Analysis of Sustainable Central Bank Communications: Revealing Authorial Signatures in Monetary Policy Statements †. *Sustainability (Switzerland)*, *17*(20), 1–15.

https://doi.org/10.3390/su17208979

Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, *14*(1), 1–15. https://doi.org/10.1038/s41598-024-54375-4

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. 1–92. http://arxiv.org/abs/2407.21783

Hamilton, R. I., & Papadopoulos, P. N. (2023). *Using_SHAP_Values_and_Machine_Learning_to_Understa (1)*. *1*, 1–12.

Hollmann, N., Müller, S., Purucker, L., & Krishnakumar, A. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(May 2024). https://doi.org/10.1038/s41586-024-08328-6

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. Bin, Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *Proceedings of Machine Learning Research*, *202*, 24950–24962.

Narendra, E. C., Dhara, F., Arsya, A., Adisty, D., Putri, Y., Informasi, S., & Komputer, F. I. (2024). *ANALISIS DAMPAK DISRUPTIF AI TERHADAP KETENAGAKERJAAN : POTENSI POSITIF DAN TANTANGAN NEGATIF ANALYSIS OF AI DISRUPTIVE IMPACT ON EMPLOYMENT : POSITIVE*. 339–350.

Of, A., Between, D., Texts, H., The, U., Language, N., & Method, P. (2025). *ANALYSIS OF DIFFERENCES BETWEEN AI AND HUMAN TEXTS USING THE NATURAL LANGUAGE PROCESSING METHOD ANALISIS PERBEDAAN TEKS BUATAN ARTIFICIAL INTELLIGENCE DAN MANUSIA MENGGUNAKAN METODE*. *10*(2), 1016–1025.

Patel, D., Parsa, S., & Johnson, A. P. (2026). A hybrid neural-symbolic approach for the longitudinal profiling of coercive control in digital investigations. *Forensic Science International: Digital Investigation*, *56*(January), 302074. https://doi.org/10.1016/j.fsidi.2026.302074

Putu, L., & Antari, S. (2025). *KESADARAN HUKUM DIGITAL DALAM PENDIDIKAN BAHASA : UPAYA MENCEGAH PLAGIARISME DAN*. *5*(1), 76–86.

Sa, D., & Gerhana, Y. A. (2025). *Deteksi Generatif Teks pada Penilaian Otomatis Tes Esai Berbahasa Indonesia Menggunakan*. *11*(2).

Tanaja, R. N., Johnny, Rafif, M. A., & Gunawan, A. A. S. (2025). Fake News Detection using Machine Learning: Integrating FakeBERT Classification, Style Analysis, and Credibility Verification. *Procedia Computer Science*, *269*, 1067–1076. https://doi.org/10.1016/j.procs.2025.09.048

Tran, V. H., Sebastian, Y., Karim, A., & Azam, S. (2024). Distinguishing Human Journalists from Artificial Storytellers Through Stylistic Fingerprints. *Computers*, *13*(12). https://doi.org/10.3390/computers13120328

Winarjo, M. A. K., & Sari, D. N. (2026). *A Systematic Comparison of SHAP and LIME for Explaining IndoBERTweet Predictions on Indonesian Toxic Content Detection*.

Journal : **Journal of Computer Science and Information Technology (JCSIT)**
Publisher : **Yayasan Nuraini Ibrahim Mandiri**
ISSN : **3031-8467**
Volume / Issue / Year/ Page : **Vol. 3 No. 4 (2026): September/ Hal 399-410**

Yatheendra, K. V., & Arabagatte, S. (2025). Cross-Linguistic Evaluation of Ai-Generated Text Detection: a Comparative Study on English and Indonesian Using Precision, Recall and F1 Score. *ShodhKosh: Journal of Visual and Performing Arts*, *6*(1), 70–75. https://doi.org/10.29121/shodhkosh.v6.i1.2025.5423