Implementation of the BiLSTM Model for Detecting AI-Generated Indonesian Text

Rafil Moehamad Alif [1)*], Syariful Alam [2)], Chandra Dewi Lestari [3)]

[1)2)3)]Teknik Informatika, Teknik, STT Wastukancana
[*)]**Correspondence author**: rafilmoehamad16@wastukancana.ac.id, Purwakarta, Indonesia
DOI: https://doi.org/10.37012/jtik.v12i1.3551

Abstract

The rapid advancement of generative Artificial Intelligence (AI) presents challenges to academic integrity due to potential misuse like plagiarism. This study develops a text detection system specifically for the Indonesian language using a Deep Learning approach with a Bidirectional Long Short-Term Memory (Bi-LSTM) architecture. The research methodology follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. A dataset comprising 5,008 text rows was compiled via web scraping from journalism platforms and academic journals indexed in SINTA 4 for human-written texts, while AI-generated counterparts were engineered using ChatGPT and Google Gemini paraphrases. Text features were extracted using a Keras Tokenizer and Embedding Layer with 64 dimensions. Evaluation of the trained Bi-LSTM model on a 30% validation split demonstrated an overall accuracy of 78.24% and a Mean Absolute Error (MAE) of 0.3295. Specifically, the model achieved a 93.77% success rate in identifying human-written texts, though it logged a lower detection rate of 62.62% for academic AI text structures. The final model was successfully deployed as a web application using Streamlit.

Keywords: *AI-Generated Text, Bi-LSTM, CRISP-DM, Deep Learning, Indonesian Language*

Abstrak

Pesatnya perkembangan teknologi *Generative Artificial Intelligence* (AI) membawa tantangan baru terhadap integritas akademik karena potensi penyalahgunaan seperti praktik plagiarisme. Penelitian ini mengembangkan sistem deteksi teks AI khusus untuk tulisan berbahasa Indonesia dengan pendekatan *Deep Learning* menggunakan arsitektur *Bidirectional Long Short-Term Memory* (Bi-LSTM). Metodologi pengembangan sistem dilakukan berdasarkan framework *Cross Industry Standard Process for Data Mining* (CRISP-DM). *Dataset* yang digunakan mencakup 5.008 data teks yang dikumpulkan melalui metode web scraping dari media jurnalistik dan abstrak jurnal ilmiah terindeks SINTA 4 untuk label teks manusia, serta hasil parafrase ChatGPT dan Google Gemini untuk label teks AI. Representasi fitur teks diekstrak menggunakan Tokenizer Keras dan Embedding Layer berdimensi 64. Hasil evaluasi model Bi-LSTM pada data validasi sebesar 30% menunjukkan tingkat akurasi sebesar 78,24% dengan nilai *Mean Absolute Error* (MAE) sebesar 0,3295. Model memiliki performa sangat baik dalam mengenali teks manusia dengan persentase keberhasilan 93,77%, namun memiliki kendala pada deteksi teks AI bernuansa akademis dengan keberhasilan sebesar 62,62%. Model akhir berhasil dideploy menjadi aplikasi *web* interaktif menggunakan *framework* Streamlit.

Kata Kunci: Bahasa Indonesia, Bi-LSTM, CRISP-DM, *Deep Learning*, Deteksi Teks AI

PENDAHULUAN

Kemajuan teknologi kecerdasan buatan (*Artificial Intelligence*/AI) khususnya pada bidang *generative AI* telah memungkinkan sistem menghasilkan konten teks otomatis secara cepat dan kontekstual yang menyerupai tulisan manusia (Mamo et al., 2024). Namun, kemudahan ini memicu fenomena penyalahgunaan fungsi utama AI, di mana teknologi yang dirancang sebagai alat bantu justru dijadikan alat utama penulisan tanpa melalui proses berpikir kritis (Bettayeb et al., 2024). Kondisi tersebut menimbulkan tantangan serius bagi dunia pendidikan karena dapat menurunkan daya kreativitas siswa serta mengancam integritas akademik lewat praktik plagiarisme terselubung (Fajt & Schiller, 2025). Teks buatan AI yang orisinal dan telah diparafrase tidak dapat diidentifikasi secara efektif oleh alat pencegah plagiarisme konvensional yang hanya mengandalkan metode pencocokan kemiripan kata (*similarity checking*) ((Fajt & Schiller, 2025).

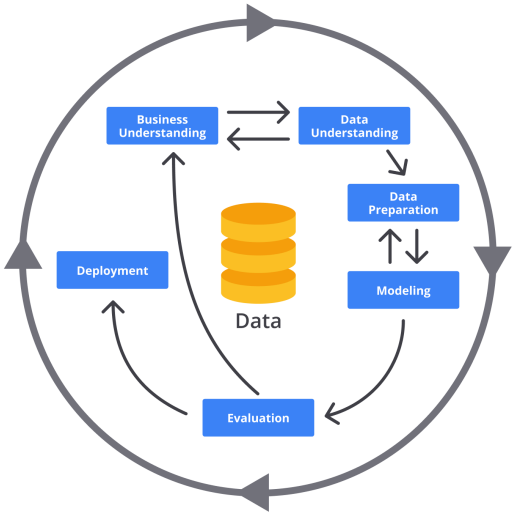
Meskipun kajian mengenai pendeteksian teks AI berkembang pesat, penelitian yang berfokus pada korpus teks berbahasa Indonesia masih relatif jarang dilakukan jika dibandingkan dengan korpus berbahasa Inggris (Pitriani et al., 2025). Guna mengatasi kesenjangan tersebut, penelitian ini bertujuan merancang, membangun, dan mengevaluasi model pembelajaran mendalam (*Deep Learning*) dengan arsitektur *Bidirectional Long Short-Term Memory* (Bi-LSTM) (Fatimah et al., 2025).

Beberapa penelitian terdahulu telah menerapkan berbagai metode untuk kebutuhan klasifikasi teks. Sebagai contoh, arsitektur Bi-LSTM telah berhasil diimplementasikan untuk kebutuhan *Named Entity Recognition* pada teks bahasa Indonesia (Pradhana et al., 2025). Untuk pemodelan deret waktu dan teks, performa *Long Short-Term Memory* konvensional juga telah teruji (Akbar et al., 2023; Owen et al., 2022). Di sisi lain, penggunaan model *pre-trained* BERT juga menunjukkan efektivitas yang tinggi dalam mendeteksi teks buatan ChatGPT (Maharani Isnainiyah et al., 2025). Penggunaan algoritma machine learning tradisional seperti Regresi Logistik (Siregar et al., 2025) maupun XGBoost dan Random Forest (Najjar et al., 2024) juga kerap dibandingkan untuk menjaga integritas akademik. Kontribusi penelitian ini adalah menyediakan model deteksi teks AI generatif versi bahasa Indonesia menggunakan representasi fitur berbasis kombinasi Tokenizer (Maulidya Prastita

Syah et al., 2025) dan Doc2Vec (Gunawan & Santoso, 2021) yang diimplementasikan langsung pada aplikasi publik berbasis Streamlit (Julyano, 2024)

METODE PENELITIAN

Metode pengerjaan sistem ini mengadopsi standar industri proses data mining yaitu *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang meliputi enam tahapan utama (Anam et al., 2026; Monesa & Jayadi, 2023). Kerangka kerja sistematis dari penelitian ini digambarkan melalui diagram alur pada Gambar 1.



Gambar 1. Alur Metode CRIPS-DM

Tahapan pelaksanaan penelitian dijabarkan secara rinci sebagai berikut:

1. Business Understanding: Mengidentifikasi masalah penyalahgunaan teks generatif AI dan merumuskan solusi berbasis *deep learning* (Jamiah Nurhakiki & Yahfizham, 2024) dengan memanfaatkan keunggulan arsitektur dua arah (Bi-LSTM) dalam menangkap konteks kalimat secara utuh (Abdillah et al., 2025).
2. Data Understanding: Proses pengumpulan data awal berbahasa Indonesia. Teks asli manusia bersumber dari artikel berita *online* (detik.com) dan dokumen abstrak artikel ilmiah dari pangkalan data Garuda Kemendikbud melalui ekstraksi menggunakan BeautifulSoup 4 (Widyastuti, 2020).
3. Data Preparation: Melakukan pembersihan data (data cleaning) berupa penyeragaman huruf kecil (*lowercase*), penghapusan karakter khusus, spasi ganda, serta tautan url web.

Tahap ini juga mencakup rekayasa data di mana teks asli dari *web* diparafrase menggunakan ChatGPT dan Google Gemini untuk memproduksi dataset kelas AI-generated (Shohifur Rizal, 2024).

4. Modelling: Mengonversi data teks bersih menjadi representasi angka menggunakan *library* Keras Tokenizer dengan batas kosakata maksimum (MAX_WORDS) sebesar 15.000 kata penting dan batasan panjang teks maksimum (MAX_LEN) sebesar 200 token kalimat. Setelah itu, arsitektur Bi-LSTM disusun menggunakan 32 unit memori yang dikombinasikan dengan L2 *Regularization* menggunakan fungsi penalti kuadrat (Kavlakogu, 2026).
5. Evaluation: Mengukur kinerja performa pengujian model menggunakan metrik evaluasi seperti Classification Report (Akurasi, Presisi, *Recall*, *F1-Score*) (Saviola & Hendrawan, 2025), Mean Absolute Error (MAE) (Suryanto & Muqtadir, 2019), serta visualisasi matriks (Confusion Matrix) (Darmawan et al., 2023).
6. Deployment: Mengintegrasikan model *deep learning* yang telah lolos uji ke dalam antarmuka aplikasi *web* interaktif berbasis *platform* Streamlit (Julyano, 2024).

HASIL DAN PEMBAHASAN

Proses pengumpulan data menghasilkan total 5.008 baris data teks yang variatif. Dataset kemudian dibagi (*data split*) secara acak dengan proporsi perbandingan nilai rasio sebesar 7:3. Rincian pembagian data disajikan pada Tabel 1.

Tabel 1. Hasil Pembagian Data (*Data Split*)

DATA	TEST	VALIDATION	TOTAL
Jumlah	3505	1503	5008
Presentase	70%	30%	100%

Model dilatih menggunakan parameter optimum dengan bantuan *optimizer* ADAM dan metode *early stopping*. Proses iterasi latihan terpantau mengalami titik konvergen pada *epoch* ke-4 dengan capaian nilai akurasi latih sebesar 78% serta tingkat kehilangan nilai validasi (*validation loss*) sebesar 54%.

Lapisan utama model adalah *Bidirectional* LSTM. Secara matematis, pemrosesan searah didasarkan pada perhitungan *forget gate* (f_t) yang ditunjukkan pada Rumus (1) (Nurashila et al., 2023)

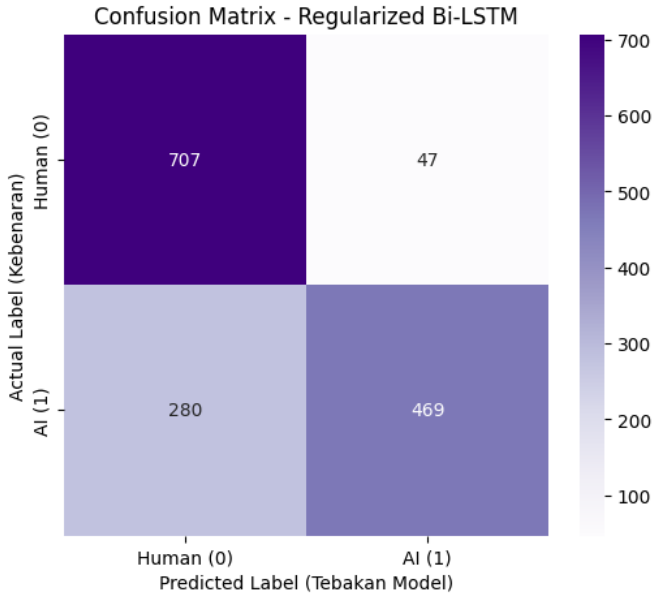
$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f \quad (1)$$

Hasil ekstraksi dari arah maju (h_t) dan arah mundur ($\overleftarrow{h_t}$) kemudian digabungkan untuk menghasilkan pemahaman konteks kalimat yang utuh sebelum diteruskan ke *Dense Layer* untuk penentuan keputusan akhir klasifikasi (Abdillah et al., 2025).

Tabel 2. Nilai Presisi dan Recall Setiap Kelas

No	Kelas Label	Precision	Recall
1	0 (Human-Written)	0.72	0.94
2	1 (AI-Generated)	0.91	0.63

Berdasarkan hasil data pada Tabel 2, model memperoleh nilai akurasi global sebesar 78,24%. Metrik Mean Absolute Error (MAE) mencatat nilai sebesar 0,3295, yang menunjukkan bahwa rata-rata bias tingkat kesalahan prediksi model dari nilai kebenaran sesungguhnya berada pada batas aman toleransi model klasifikasi (Suryanto et al., 2019).



Gambar 2. Hasil *Confusion Matrix*

Analisis performa spesifik didapatkan melalui pemetaan *Confusion Matrix* dengan rincian *True Negative* (TN) = 707 data, *False Positive* (FP) = 47 data, *False Negative* (FN) = 280 data, dan *True Positive* (TP) = 469 data. Berdasarkan nilai tersebut, diperoleh dua poin analisis utama:

1. Model menunjukkan kemampuan yang sangat tinggi dalam mengenali teks asli buatan manusia dengan persentase akurasi kelas mencapai 93,77%

$$\frac{707}{707+47} \boldsymbol{X}\ 100\% = 93.77\%$$

2. Model memiliki keterbatasan dalam mendeteksi teks buatan AI dengan persentase akurasi kelas sebesar 62,62%.

$$\frac{469}{469+280} \boldsymbol{X}\ 100\% = 62.62\%$$

Kesalahan klasifikasi *False Negative* (teks AI terdeteksi sebagai manusia) dipengaruhi oleh karakteristik generator AI saat melakukan parafrase pada teks artikel ilmiah (Elkhatat et al., 2023). Generator AI cenderung mengadopsi pilihan kata baku dan struktur kalimat formal yang memiliki pola distribusi linguistik beririsan langsung dengan gaya penulisan akademik asli milik manusia (Siregar et al., 2025). Model akhir yang telah terbentuk berhasil diimplementasikan ke dalam platform web *Streamlit* yang memuat menu utama berupa halaman panduan, antarmuka deteksi teks, serta halaman informasi sistem (Julyano, 2024).

KESIMPULAN DAN REKOMENDASI

Penelitian ini telah berhasil mengembangkan aplikasi deteksi teks *AI-generated* khusus untuk tulisan berbahasa Indonesia dengan arsitektur *Deep Learning* berbasis *Bidirectional Long Short-Term Memory* (Bi-LSTM) (Fatimah et al., 2025). Berdasarkan hasil evaluasi pengujian pada 1.503 data validasi, model mampu beroperasi secara konsisten dengan perolehan akurasi menyeluruh sebesar 78,24% dan nilai kesalahan absolut (MAE) sebesar 0,3295. Secara mendalam, model ini sangat andal dalam menyaring tulisan asli karya manusia dengan tingkat keberhasilan sebesar 93,77%, namun masih memiliki kendala dalam mengidentifikasi teks buatan AI yang menggunakan gaya bahasa akademik formal dengan tingkat deteksi sebesar 62,62%. Permasalahan riset mengenai deteksi plagiarisme terselubung teks AI bahasa Indonesia telah berhasil dijawab melalui pembuktian efisiensi model Bi-LSTM ini. Untuk pengembangan berikutnya, direkomendasikan penggunaan teknik *transfer learning* menggunakan model replika lokal yang lebih masif seperti

IndoBERT guna meningkatkan sensitivitas pengenalan struktur teks AI akademis secara *real-time* (Pitriani et al., 2025)**.**

REFERENSI

Abdillah, F., Hadiana, A. I., & Melina, M. (2025). Prediksi Curah Hujan Menggunakan Metode Bi-LSTM dan GRU Berbasis Data Iklim. *Jurnal Algoritma*, *22*(2), 12–21. https://doi.org/10.33364/algoritma/v.22-2.2305

Akbar, R., Santoso, R., & Warsito, B. (2023). Prediksi Tingkat Temperatur Kota Semarang Menggunakan Metode Long Short-Term Memory (LSTM). *Jurnal Gaussian*, *11*(4), 572–579. https://doi.org/10.14710/j.gauss.11.4.572-579

Anam, F. S., Muttaqin, M. R., & Ramadhan, Y. R. (2026). *Klasifikasi Penyakit Pada Daun dan Buah Jambu Menggunakan Convolutional Neural Network* (Vol. 7, Number 1).

Bettayeb, A. M., Abu Talib, M., Sobhe Altayasinah, A. Z., & Dakalbab, F. (2024). Exploring the impact of ChatGPT: conversational AI in education. In *Frontiers in Education* (Vol. 9). Frontiers Media SA. https://doi.org/10.3389/feduc.2024.1379796

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1). https://doi.org/10.1007/s40979-023-00140-5

Fajt, B., & Schiller, E. (2025). ChatGPT in Academia: University Students' Attitudes Towards the use of ChatGPT and Plagiarism. *Journal of Academic Ethics*, *23*(3), 1363–1382. https://doi.org/10.1007/s10805-025-09603-5

Gunawan, K. I., & Santoso, J. (2021). Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia. *Journal of Information System,Graphics, Hospitality and Technology*, *3*(01), 29–38. https://doi.org/10.37823/insight.v3i01.126

Julyano, M. B. (2024). Desain dan Implementasi Website Harvest Lens Prediksi Harga Beras Menggunakan Framework Streamlit. *E-Proceeding of Engineering*, *11*, 2856–2860.

Maharani Isnainiyah, P., Sri Kusuma Aditya, C., & Rizki Chandranegara, D. (2025). Penerapan Model Pre-Trained BERT dalam Mendeteksi Teks Buatan ChatGPT. *REPOSITOR*, *7*(3), 393–400.

Mamo, Y., Crompton, H., Burke, D., & Nickel, C. (2024). Higher Education Faculty Perceptions of ChatGPT and the Influencing Factors: A Sentiment Analysis of X. *TechTrends*, *68*(3), 520–534. https://doi.org/10.1007/s11528-024-00954-1

Maulidya Prastita Syah, Ajeng Puspa Wardani, Mohammad Idhom, & Trimono. (2025). Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks. *Data Sciences Indonesia (DSI)*, *5*(1), 47–59. https://doi.org/10.47709/dsi.v5i1.6021

Monesa, F., & Jayadi, R. (2023). Sentiment Analysis On Investment Topic In Indonesia Using Machine Learning Algorithms Approach. *Journal of Theoretical and Applied Information Technology*, *15*(13). www.jatit.org

Najjar, A. A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2024). *Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity*.

Nurashila, S. S., Hamami, F., & Kusumasari, T. F. (2023). Perbandingan Kinerja Algoritma Recurrent Neural Network (RNN) Dan Long Short-Term Memory (LSTM): Studi Kasus Prediksi Kemacetan Lalu Lintas Jaringan Pt Xyz. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *8*(3), 864–877. https://doi.org/10.29100/jipi.v8i3.3961

Owen, M., Vincent, V., Br Ambarita, R., & Indra, E. (2022). Implementasi Metode Long Short Term Memory Untuk Memprediksi Pergerakan Nilai Harga Emas. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *5*(1), 96. https://doi.org/10.37600/tekinkom.v5i1.507

Pitriani, Maylawati, D. S., & Gerhana, Y. A. (2025). Deteksi Generatif Teks pada Penilaian Otomatis Tes Esai Berbahasa Indonesia Menggunakan IndoBERT. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *11*.

Siregar, A. A., Indra, Z., Arnita, Taufik, I., & Niska, D. Y. (2025). Pengembangan Aplikasi Deteksi Teks Menggunakan Algoritma Regresi Logistik. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*.

Suryanto, A. A., Muqtadir, A., & Artikel, S. (2019). *Penerapan Metode Mean Absolute Error (Mae) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi Info Artikel : ABSTRAK*. (1), 11.