

Prediksi Gender Berdasarkan Nama Menggunakan Kombinasi Model IndoBERT, *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM)

Abi Mas'ud ^{1,*}, Bambang Krismono Triwijoyo ¹ dan Dadang Priyanto ¹

¹ Magister Ilmu Komputer, Universitas Bumigora, Indonesia

* Korespondensi: abimsama01@gmail.com

Sitasi: Mas'ud, A.; Triwijoyo, K, B.; Priyanto, D. (2025). Prediksi Gender Berdasarkan Nama Menggunakan Kombinasi Model IndoBERT, *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM). JTIM: Jurnal Teknologi Informasi Dan Multimedia, 7(3), 448-460. <https://doi.org/10.35746/jtim.v7i3.736>

Diterima: 08-05-2025

Direvisi: 10-06-2025

Disetujui: 16-06-2025



Copyright: © 2025 oleh para penulis. Karya ini dilisensikan di bawah Creative Commons Attribution-ShareAlike 4.0 International License. (<https://creativecommons.org/licenses/by-sa/4.0/>).

Abstract: This study proposes a name-based gender prediction model in the Indonesian language by combining the architectures of Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM). The non-standardized and diverse structure of Indonesian names presents a significant challenge for text-based gender classification tasks. To address this, a hybrid approach was developed to leverage the contextual representation power of IndoBERT, the local pattern extraction capability of CNN, and the sequential dependency modeling strength of BiLSTM. The dataset consists of 4,796 student names from Universitas Bumigora, collected between 2018 and 2023. The preprocessing steps include lowercasing, punctuation removal, label encoding, and train-test splitting. Evaluation results based on accuracy, precision, recall, and F1-score indicate that the IndoBERT-CNN-BiLSTM model achieved the best performance, with an accuracy of 90.94%, F1-score of 91.03%, and training stability without signs of overfitting. This model demonstrates high effectiveness in name-based gender classification and holds strong potential for applications such as population information systems, service personalization, and name-based demographic analysis.

Keywords: Gender prediction, IndoBERT, CNN, BiLSTM, text classification and deep learning

Abstrak: Penelitian ini mengusulkan model prediksi gender berbasis nama dalam bahasa Indonesia dengan menggabungkan arsitektur *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT), *Convolutional Neural Network* (CNN), dan *Bidirectional Long Short-Term Memory* (BiLSTM). Kompleksitas struktur nama dalam bahasa Indonesia yang tidak terstandarisasi menjadi tantangan dalam tugas klasifikasi jenis kelamin berbasis teks. Untuk mengatasi hal tersebut, pendekatan hibrida dikembangkan guna memanfaatkan kekuatan representasi kontekstual IndoBERT, kemampuan ekstraksi pola lokal CNN, serta pemahaman dependensi sekuensial BiLSTM. Dataset yang digunakan terdiri dari 4.796 nama mahasiswa Universitas Bumigora dari tahun 2018 hingga 2023. Proses preprocessing meliputi *lowercasing*, penghapusan tanda baca, *label encoding*, serta *train-test split*. Hasil evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* menunjukkan bahwa model IndoBERT-CNN-BiLSTM mencapai performa terbaik dengan akurasi sebesar 90,94%, F1-score 91,03%, serta kestabilan dalam pelatihan tanpa indikasi *overfitting*. Model ini menunjukkan kapabilitas yang tinggi dalam klasifikasi gender berbasis nama dan berpotensi diterapkan dalam berbagai aplikasi seperti sistem informasi kependudukan, personalisasi layanan, dan analisis demografi berbasis nama.

Kata kunci: Prediksi gender, IndoBERT, CNN, BiLSTM, klasifikasi teks dan Deep Learning.

1. Pendahuluan

Prediksi *gender* berdasarkan nama merupakan salah satu bidang kajian penting dalam pengolahan bahasa alami (NLP) dan pembelajaran mesin [1]. Informasi *gender* sering dibutuhkan untuk meningkatkan ketepatan dalam berbagai sistem seperti pemasaran, keamanan, analisis data pengguna, dan interaksi berbasis mesin [2]. Berbagai metode telah dikembangkan untuk prediksi *gender*, mulai dari pendekatan menggunakan daftar nama, aturan linguistik, hingga teknik pembelajaran mendalam berbasis *neural network*. Namun, masih terdapat tantangan dalam memprediksi *gender* secara akurat, terutama untuk nama-nama yang berasal dari bahasa dengan struktur dan pola yang bervariasi [3].

Pada bahasa Indonesia, karakteristik nama yang beragam, tidak terstandarisasi, dan sering kali dipengaruhi oleh budaya lokal membuat prediksi *gender* menjadi lebih kompleks [4]. Untuk itu, diperlukan model yang mampu memahami makna kontekstual dari nama secara lebih mendalam [5]. *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT), sebagai model berbasis *transformer* yang dilatih khusus untuk bahasa Indonesia, memiliki kemampuan representasi konteks yang kuat [6]. Sementara itu, *Convolutional Neural Network* (CNN) efektif dalam mengekstraksi pola lokal pada data sekuensial [7], dan *Bidirectional Long Short-Term Memory* (BiLSTM) mampu menangkap hubungan jangka panjang dari kedua arah data masukan [7].

Penelitian ini mengusulkan model gabungan IndoBERT, CNN dan BiLSTM untuk memprediksi *gender* berdasarkan nama. Dengan mengombinasikan keunggulan ketiga model tersebut, diharapkan mampu meningkatkan ketepatan prediksi *gender* pada nama-nama berbahasa Indonesia, serta mengatasi tantangan yang muncul akibat variasi struktur nama.

Metode gabungan IndoBERT, CNN, dan BiLSTM digunakan karena ketiganya memiliki kelebihan yang saling melengkapi. IndoBERT memahami konteks kata dalam bahasa Indonesia dengan baik, CNN menangkap pola-pola lokal seperti awalan atau akhiran pada nama, dan BiLSTM memahami urutan nama dari dua arah. Dengan menggabungkan ketiganya, model dapat mengenali makna, pola, dan struktur nama secara lebih menyeluruh, sehingga lebih tepat dalam memprediksi *gender* berdasarkan nama.

Tujuan utama dari penelitian ini adalah mengembangkan sebuah model prediksi jenis kelamin berdasarkan nama dengan menggunakan metode gabungan IndoBERT, CNN dan BiLSTM. Penelitian ini bertujuan untuk menginvestigasi pola-pola sekuensial yang terdapat dalam nama serta karakteristik-karakteristik lokal yang terkait dengan *gender*, sekaligus memperkenalkan teknik encoding tingkat karakter dan kata menggunakan embedding dari IndoBERT. Fokus penelitian diarahkan pada pengembangan model prediksi *gender* yang memiliki akurasi tinggi dan dapat diandalkan ketika diterapkan pada berbagai nama. Selain itu, penelitian ini juga bertujuan untuk mengatasi keterbatasan metode-metode tradisional dalam memahami ketergantungan jangka panjang dan pola-pola lokal dalam data sekuensial.

Penelitian ini juga akan mengevaluasi kinerja model gabungan melalui serangkaian eksperimen untuk menguji metrik performa seperti *accuracy*, *precision*, *recall*, dan *F1-score*, serta membandingkannya dengan metode prediksi lainnya [8]. Dengan demikian, hasil dari penelitian ini diharapkan mampu memberikan kontribusi signifikan dalam pengembangan aplikasi berbasis Bahasa Indonesia yang lebih cerdas, khususnya dalam sektor-sektor yang membutuhkan analisis *gender* berbasis data nama, seperti sistem pendaftaran, analisis demografi, dan layanan personalisasi [7].

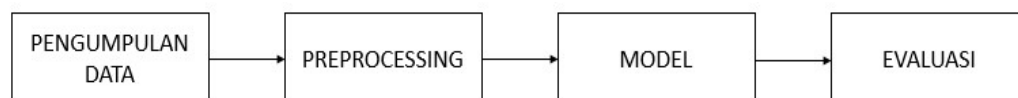
2. Penelitian Terdahulu

Penelitian tentang prediksi *gender* berdasarkan nama telah banyak dilakukan dalam berbagai bahasa. [2] Salah satunya menggunakan dataset UIT-ViNames berisi 26.850 nama Vietnam, dan model LSTM berhasil memberikan hasil terbaik dengan *F1-score* sebesar 95,89%. Penelitian lain [9] menggunakan 100.787 nama dari Brasil dan lima model, menunjukkan bahwa model berbasis *recurrent* seperti BiLSTM memberikan akurasi tertinggi sebesar 95,89%, mengungguli model *feedforward* seperti MLP dan CNN. Penelitian [10] pada nama Bengali menggunakan 5.748 data menunjukkan bahwa *Neural Network* memberikan hasil terbaik dengan akurasi 73,04%, mengungguli model seperti KNN, *Random Forest*, dan *Naive Bayes*. Penelitian lain [11] dengan 84.899 nama multinasional menunjukkan bahwa model *SimpleText* berbasis *deep learning* memberikan akurasi tertinggi sebesar 94,67%, sedikit lebih baik dibandingkan Linear SVC dan model lainnya. Di Indonesia [4] penelitian menggunakan nama berbahasa Indonesia dari data pemilih dan mahasiswa menunjukkan bahwa model LSTM mencapai akurasi pengujian 96%, lebih tinggi dibandingkan *Naive Bayes* yang hanya 93%, meskipun akurasi pelatihan LSTM sedikit lebih rendah.

Penelitian terdahulu umumnya masih terbatas pada penerapan satu jenis model pembelajaran mendalam, seperti LSTM atau CNN secara terpisah, yang belum sepenuhnya mampu menangkap kompleksitas struktur nama dalam bahasa Indonesia yang cenderung tidak terstandarisasi. Selain itu, sebagian besar studi dilakukan pada dataset berbahasa asing, sehingga kurang merepresentasikan karakteristik linguistik dan budaya nama-nama Indonesia. Oleh karena itu, penelitian ini menawarkan pendekatan baru yang lebih komprehensif melalui kombinasi arsitektur IndoBERT, CNN, dan BiLSTM, guna mengintegrasikan pemahaman kontekstual, pola lokal, dan dependensi sekuensial untuk meningkatkan akurasi prediksi gender berbasis nama secara signifikan.

3. Metodologi Penelitian

Tahapan penelitian menjelaskan tahapan – tahapan yang dilakukan dalam penelitian ini, adapun tahapan pada penelitian ini dimulai dari pengumpulan data, *preprocessing*, model, dan evaluasi. Alur pada penelitian ini dapat dilihat pada Gambar 1 sebagai berikut:



Gambar 1. Alur Penelitian

Gambar di atas menggambarkan alur penelitian prediksi gender berbasis nama yang terdiri dari empat tahapan utama [4]. Tahap pertama adalah pengumpulan data, yaitu menghimpun data nama lengkap mahasiswa beserta jenis kelamin sebagai label. Selanjutnya, dilakukan *preprocessing* data yang mencakup normalisasi teks (seperti konversi huruf menjadi huruf kecil dan penghapusan tanda baca), pelabelan data dalam bentuk numerik, serta proses *text encoding* untuk mengubah data teks menjadi representasi vektor. Tahap ketiga adalah penerapan model, yaitu implementasi arsitektur gabungan IndoBERT, CNN, dan BiLSTM untuk melakukan prediksi jenis kelamin berdasarkan nama. Tahap akhir adalah evaluasi kinerja model, yang dilakukan dengan mengukur metrik performa seperti *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai efektivitas model dalam melakukan klasifikasi.

3.1. Pengumpulan Data

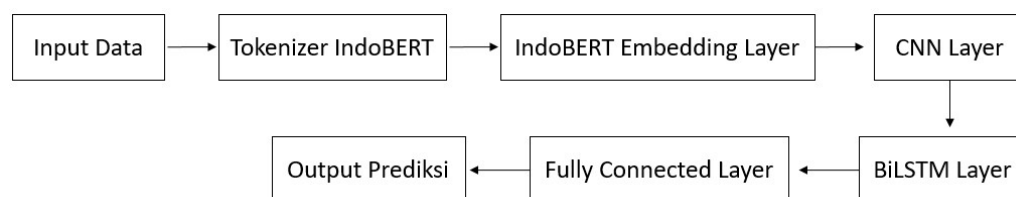
Dalam penelitian ini, data yang digunakan berasal dari Universitas Bumigora dan diperoleh melalui pengumpulan nama lengkap mahasiswa dari berbagai program studi. Data mencakup seluruh mahasiswa aktif dari angkatan tahun 2018 hingga tahun 2023, sehingga mewakili populasi lintas jurusan dan tingkat akademik. Jumlah total data yang berhasil dikumpulkan sebanyak 4.796 entri, yang masing-masing memuat informasi nama lengkap dan jenis kelamin mahasiswa. Dari total 4.796 mahasiswa tersebut, terdapat 3.060 mahasiswa laki-laki, yang mewakili sekitar 64% dari keseluruhan data. Sementara itu, terdapat 1.736 mahasiswa perempuan, yang mencakup 36% dari total populasi mahasiswa dalam rentang tahun tersebut. Proporsi ini mencerminkan distribusi *gender* aktual di lingkungan Universitas Bumigora selama periode yang diamati, dan menjadi acuan dalam proses pelabelan data serta pengembangan model prediksi *gender* berbasis nama.

3.2. Preprocessing Data

Tahapan *preprocessing* data dalam penelitian ini terdiri dari lima langkah utama yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pelatihan model [12]. Pertama, dilakukan proses *lowercasing*, yaitu mengubah seluruh huruf dalam nama menjadi huruf kecil guna menyamakan representasi kata dan menghindari perbedaan makna antara huruf besar dan kecil. Kedua, *punctuation removal* diterapkan untuk menghapus tanda baca yang tidak relevan, sehingga model dapat lebih fokus pada informasi karakter yang bermakna. Ketiga, dilakukan identifikasi kata atau karakter unik untuk menentukan jumlah fitur maksimum (*max features*), yang bertujuan menyeimbangkan antara kompleksitas model dan kapasitas generalisasi. Keempat, label *gender* yang semula berbentuk teks dikonversi menjadi bentuk numerik melalui proses *label encoding*, sehingga dapat dikenali dan diproses oleh algoritma *machine learning* dan *deep learning* secara efisien. Terakhir, data dibagi menjadi data latih dan data uji menggunakan metode *train-test split*, untuk mengevaluasi performa model secara objektif dan memastikan kemampuannya dalam melakukan generalisasi pada data yang belum pernah dilihat sebelumnya [13].

3.3. Model – Hybrid IndoBERT, CNN dan BiLSTM

Setelah melalui tahap *preprocessing* data, langkah selanjutnya adalah penerapan model untuk proses pelatihan dan pengujian [14]. Pada tahap ini, data yang telah dipersiapkan dianalisis menggunakan pendekatan arsitektur gabungan yang terdiri dari *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT), *Convolutional Neural Network* (CNN), dan *Bidirectional Long Short-Term Memory* (BiLSTM). Model ini dirancang untuk mengintegrasikan kekuatan representasi kontekstual IndoBERT, kemampuan ekstraksi pola lokal dari CNN, serta pemahaman sekuensial jangka panjang dari BiLSTM [3]. Arsitektur gabungan ini diimplementasikan guna meningkatkan efektivitas dalam klasifikasi teks berbasis nama [3]. Ilustrasi arsitektur secara umum dapat dilihat pada Gambar 2.



Gambar 2. Alur Model Prediksi Gabungan IndoBERT-CNN-BiLSTM

Model gabungan yang terdiri dari IndoBERT-CNN-BiLSTM dirancang untuk meningkatkan pemahaman teks berbahasa Indonesia dengan memanfaatkan keunggulan masing-masing arsitektur. IndoBERT digunakan sebagai *feature extractor* untuk menghasilkan representasi kontekstual teks, dilatih khusus pada korpus Bahasa Indonesia. Representasi ini kemudian diproses oleh CNN untuk mengekstraksi pola lokal, seperti n-gram, yang relevan dalam struktur kata. Selanjutnya, hasil ekstraksi diteruskan ke BiLSTM untuk menangkap dependensi sekuensial dalam teks dari dua arah, memperkaya pemahaman konteks dan urutan kata. Gabungan ini memungkinkan model untuk mengidentifikasi makna semantik, pola lokal, dan urutan informasi secara lebih efektif, meningkatkan akurasi dalam tugas-tugas pemrosesan bahasa alami, seperti klasifikasi teks dan prediksi berdasarkan nama [15].

Self-Attention (IndoBERT):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Merupakan mekanisme utama dalam arsitektur Transformer yang digunakan untuk memperoleh representasi kontekstual dari setiap token. Dalam proses ini, Query (Q) dibandingkan dengan Key (K) untuk menghasilkan skor perhatian yang menunjukkan seberapa relevan suatu token terhadap token lainnya. Skor tersebut dinormalisasi dengan membaginya dengan akar dari dimensi Key ($\sqrt{d_k}$) dan kemudian diterapkan fungsi softmax untuk mendapatkan bobot perhatian. Bobot ini selanjutnya digunakan untuk mengalikan Value (V), sehingga menghasilkan representasi akhir dari token yang telah mempertimbangkan konteks global dari seluruh urutan input.

Convolution (CNN):

$$h(i) = ReLu(W_{cnn} * E_{bert}[i:i+k-1] + b) \quad (2)$$

Menjelaskan cara model mengambil bagian kecil dari hasil BERT, lalu memprosesnya dengan filter konvolusi W_{cnn} untuk menangkap pola-pola penting dalam teks. Potongan embedding dari posisi i sampai $i+k-1$ dihitung dengan filter tersebut, ditambahkan bias b , lalu dilewatkan ke fungsi ReLU agar hanya nilai positif yang diteruskan. Hasil akhirnya, yaitu $h(i)$, adalah vektor yang berisi informasi penting dari bagian teks itu, dan bisa digunakan untuk langkah selanjutnya seperti pooling atau klasifikasi.

BiLSTM Processing:

$$\begin{aligned} \vec{h}_t, \overleftarrow{h}_t &= LSTM(h(i)) \\ h_t &= [\vec{h}_t; \overleftarrow{h}_t] \end{aligned} \quad (3)$$

Rumus tersebut menunjukkan bahwa hasil dari CNN, yaitu $h(i)$, diproses oleh LSTM dua arah (maju dan mundur). Output dari LSTM maju adalah \vec{h}_t , dan dari LSTM mundur adalah \overleftarrow{h}_t . Kedua hasil ini kemudian digabungkan menjadi satu vektor h_t . Dengan cara ini, model bisa memahami konteks kata dari arah depan dan belakang secara bersamaan, sehingga makna teks bisa ditangkap lebih lengkap.

$$\hat{y} = softmax(W_{fc}h_t + b_{fc}) \quad (4)$$

Rumus tersebut menunjukkan proses akhir untuk membuat prediksi. Vektor fitur h_t yang berisi informasi dari teks dimasukkan ke dalam lapisan fully connected dengan bobot W_{fc} dan bias b_{fc} . Hasilnya kemudian dihitung menggunakan fungsi softmax agar menjadi nilai probabilitas. Output \hat{y} adalah prediksi model, yaitu probabilitas untuk setiap kelas, dan kelas dengan nilai tertinggi dianggap sebagai hasil prediksi akhir.

3.4. Evaluasi

Tahapan evaluasi dilakukan untuk pengujian performa model dengan menggunakan *confusion matrix*. Pada tahap ini, akan menghasilkan metrik evaluasi berupa *accuracy*, *precision*, *recall*, dan *f1-score*. Untuk mengukur metrik evaluasi pada penelitian ini bertujuan untuk mencari nilai *accuracy*, *precision*, *recall*, dan *f1-score* [16].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

3.5. Hasil

Hasil dari penelitian ini bertujuan untuk membuktikan bahwa model yang dikembangkan telah memenuhi ekspektasi dan dapat diimplementasikan secara efektif. Evaluasi model dilakukan dengan menggunakan *confusion matrix*, yang menghasilkan beberapa metrik evaluasi, antara lain *accuracy*, *precision*, *recall*, dan *f1-score*. Proses evaluasi ini dilakukan dengan menerapkan model gabungan IndoBERT-CNN-BiLSTM. Untuk memperkuat analisis, penelitian ini juga membandingkan performa model utama tersebut dengan dua model pembanding sebagai *benchmark*, yaitu model IndoBERT-CNN dan IndoBERT-BiLSTM. Berikut hasil perbandingan kinerja ketiga model tersebut dapat dilihat pada Tabel berikut.

3.5.1. Pengumpulan Data

Dalam penelitian ini, data yang digunakan diperoleh dari Universitas Bumigora, berupa nama lengkap mahasiswa yang mencakup seluruh jurusan dari angkatan tahun 2018 hingga 2023. Total jumlah mahasiswa yang terkumpul adalah sebanyak 4.796 orang, dengan rincian 3.060 mahasiswa (64%) berjenis kelamin laki-laki dan 1.736 mahasiswa (36%) berjenis kelamin perempuan. Data ini menjadi dasar dalam proses pelatihan dan evaluasi model untuk prediksi *gender* berdasarkan nama.

Tabel 1. Hasil Pengumpulan Data Mahasiswa Universita Bumigora.

	Nama Mahasiswa	Jenis Kelamin
0	YUDISTIRA	L
1	MUHAMMAD DODI ZULIANSYAH	L
2	MUJI BAITUL BAHILI	P
...
4795	M ALFATIH MURAD SUJIWO SETO P	L
4796	FELIX HONGGADI PUTRA	L
4797	JAYA ANGGRAWAN	L

3.5.2. Preprocessing Data

Sebelum data digunakan dalam proses pelatihan model, dilakukan serangkaian tahapan *preprocessing* untuk memastikan kualitas dan konsistensi data yang optimal [2]. Tahapan *preprocessing* yang diterapkan dalam penelitian ini adalah sebagai berikut:

1). Lowercasing

Lowercasing dilakukan dengan mengubah semua huruf menjadi huruf kecil. Tujuannya untuk menyamakan format penulisan agar sistem tidak membedakan kata atau nama yang sama namun ditulis dengan kapitalisasi berbeda, seperti "Yudistira", "YUDISTIRA", dan "yudistira". Langkah ini membantu meningkatkan konsistensi data dan efisiensi pemrosesan teks. Hasilnya ditampilkan pada Tabel 2.

Tabel 2. Proses Mengubah Semua Huruf Dalam Teks Menjadi Huruf Kecil.

	Nama Mahasiswa	Jenis Kelamin
0	yudistira	L
1	muhammad dodi zuliansyah	L
2	muji baitul baheli	P
3	made abhi dwiva melia gana	L
4	sumaidi	L

2). Punctuation Removal

Proses ini menghapus seluruh tanda baca dalam data teks, seperti titik, koma, tanda seru, dan tanda tanya. Tujuannya adalah untuk membersihkan teks dari karakter non-alfabetik yang tidak memberikan kontribusi berarti terhadap makna. Dengan menghilangkan tanda baca, model dapat lebih fokus pada informasi yang relevan, sehingga meningkatkan kualitas data dan kinerja model dalam analisis atau klasifikasi teks. Hasilnya ditampilkan pada Tabel 3.

Tabel 3. Proses Penghapusan Tanda Baca.

	Nama Mahasiswa	Has Punctuation	Punctuation Info
36	helkaturrabi'ah	True	{':': 1}
58	rizky abd.ashary	True	{':': 1}
70	m. azwar hidayat	True	{':': 1}
...
4740	muh rifa'i	True	{':': 1}
4784	m.apriza ariantom. rijal	True	{':': 1}
4794	m. rijal	True	{':': 1}

3). Check Unique Words for Max Features

Proses ini menghitung jumlah kata unik dalam data teks untuk mengetahui tingkat variasi kata. Informasi ini digunakan sebagai dasar dalam menentukan batas jumlah fitur saat teks diubah menjadi representasi numerik, seperti pada proses tokenisasi. Penetapan batas ini penting untuk menjaga efisiensi komputasi dan menghindari *overfitting* yang dapat menurunkan performa model. Hasilnya ditampilkan pada Tabel 4.

Tabel 4. Proses Menghitung Jumlah Kata Unik Dalam Data Teks.

Tanda (')	Tanda (.)	Tanda (-)
50	204	18

4). Label Encoding

Proses ini mengubah data kategorikal menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*. Dalam penelitian ini, kategori jenis kelamin 'L' (Laki-laki) dan 'P' (Perempuan) dikodekan menjadi 0 dan 1, lalu disimpan dalam kolom bernama label. Langkah ini penting karena sebagian besar algoritma hanya menerima input numerik, sehingga konversi ini menjadi tahap awal yang krusial dalam pelatihan model. Hasilnya ditampilkan pada Tabel 5.

Tabel 5. Proses Mengonversi Nilai Kategori Menjadi Angka.

Jenis Kelamin	Keterangan	Label
L	Laki-laki	0
P	Perempuan	1

5). Train Test Split

Dataset dibagi menjadi dua bagian: data latih untuk membangun model dan data uji untuk mengevaluasinya. Nama mahasiswa digunakan sebagai fitur (*input*), sedangkan jenis kelamin sebagai label (*output*). Sebanyak 80% data digunakan untuk pelatihan dan 20% untuk pengujian. Parameter *random_state=42* digunakan agar pembagian data konsisten, dan *stratify* diterapkan untuk menjaga proporsi kelas tetap seimbang pada kedua bagian, sehingga evaluasi model lebih adil dan representatif.

3.5.3. Model – Hybrid IndoBERT-CNN-BiLSTM

Setelah melakukan proses *preprocessing* data, tahap selanjutnya adalah melakukan penerapan model. Pada tahap penerapan model, data pelatihan dan data pengujian di proses dan dianalisis menggunakan algoritma atau model yang diterapkan yaitu pendekatan gabungan *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT), *Convolutional Neural Network* (CNN), dan *Bidirectional Long Short-Term Memory* (BiLSTM). Berikut adalah hasil menggunakan pendekatan gabungan IndoBERT-CNN-BiLSTM.

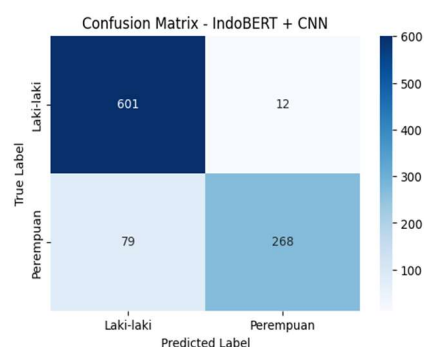
Tabel 6. Performa model *Hybrid* IndoBERT-CNN-BiLSTM.

	precision	recall	f1-score	support
0	0.9598	0.8956	0.9266	613
1	0.8351	0.9337	0.8816	347
accuracy			0.9094	960
macro avg	0.8974	0.9147	0.9041	960
weighted avg	0.9147	0.9094	0.9103	960

Berdasarkan hasil evaluasi pada Tabel 6, model gabungan IndoBERT-CNN-BiLSTM menunjukkan kinerja yang optimal dengan *accuracy* sebesar 90,94%. Untuk kelas 0, diperoleh *precision* sebesar 0,9598, *recall* 0,8956, dan *f1-score* 0,9266, menandakan kemampuan model yang tinggi dalam mengidentifikasi kelas ini. Pada kelas 1, *precision* sebesar 0,8351 dan *recall* 0,9337 menghasilkan *f1-score* 0,8816, yang mencerminkan sensitivitas tinggi meskipun presisi sedikit lebih rendah. Nilai rata-rata *macro* untuk *precision*, *recall*, dan *f1-score* masing-masing adalah 0,8974, 0,9147, dan 0,9041, sedangkan rata-rata tertimbangnya mencapai 0,9147, 0,9094, dan 0,9103. Secara keseluruhan, model mampu melakukan klasifikasi secara efektif dan seimbang, meskipun terdapat ketimpangan distribusi data antar kelas.

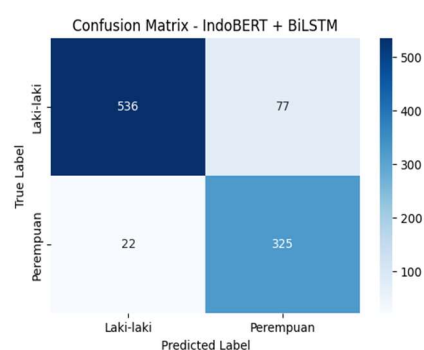
3.5.4. Evaluasi

Pada tahap evaluasi, dilakukan perbandingan kinerja tiga arsitektur model *hybrid*: IndoBERT-CNN, IndoBERT-BiLSTM, dan IndoBERT-CNN-BiLSTM. Tujuannya adalah untuk menentukan model paling efektif dalam klasifikasi jenis kelamin berbasis nama. Evaluasi menggunakan *metrix accuracy*, *precision*, *recall*, dan *f1-score* yang dihitung berdasarkan *confusion matrix*. *Confusion matrix* memberikan gambaran tabular atas prediksi model dibandingkan dengan label aktual, sehingga memungkinkan analisis menyeluruh terhadap kemampuan klasifikasi dan generalisasi masing-masing model terhadap data uji.



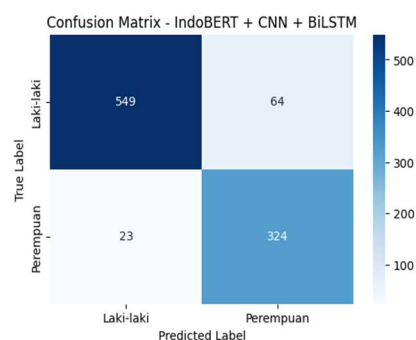
Gambar 3. *Confusion Matrix* IndoBERT-CNN

Pada gambar 3 menunjukkan hasil evaluasi model IndoBERT yang dikombinasikan dengan CNN untuk klasifikasi *gender* berbasis nama dalam bahasa Indonesia. Model berhasil mengklasifikasikan 601 nama laki-laki dan 268 nama perempuan secara benar, dengan tingkat *accuracy* sebesar 90,5%. *Precision* untuk kelas perempuan mencapai 95,7%, *recall* 77,2%, dan *F1-score* sebesar 85,5%. Hasil ini mengindikasikan bahwa model memiliki kinerja tinggi, khususnya dalam mengenali nama laki-laki, meskipun performa pada kelas perempuan masih perlu ditingkatkan. Perbedaan ini dapat disebabkan oleh variasi morfologis pada nama perempuan atau ketidakseimbangan distribusi data antar kelas.



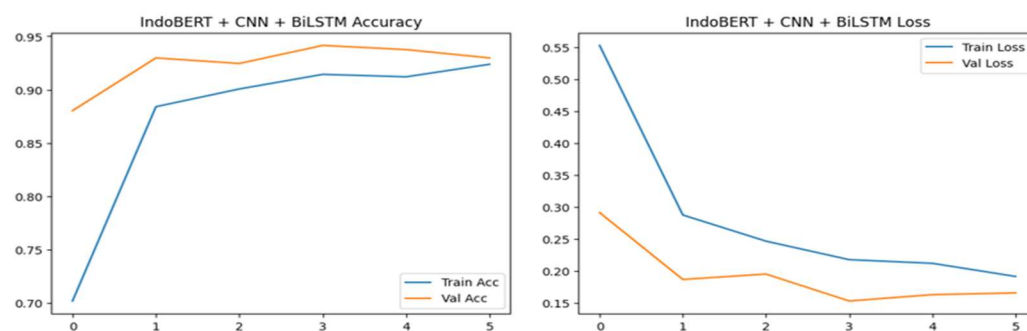
Gambar 4. *Confusion Matrix* IndoBERT-BiLSTM

Pada gambar 4, menunjukkan hasil evaluasi model IndoBERT yang dikombinasikan dengan BiLSTM untuk klasifikasi *gender* berdasarkan nama dalam bahasa Indonesia. Model berhasil mengklasifikasikan 536 nama laki-laki dan 325 nama perempuan secara tepat, dengan *accuracy* keseluruhan sebesar 89,6%. *Precision* untuk kelas perempuan mencapai 80,8%, *recall* 93,7%, dan *F1-score* 86,8%. Hasil ini menunjukkan bahwa model cukup seimbang dalam mengenali kedua kelas, dengan kinerja yang sangat baik dalam mengidentifikasi nama perempuan, berkat kemampuan BiLSTM dalam menangkap informasi berurutan dari nama.



Gambar 5. *Confusion Matrix* IndoBERT-CNN-BiLSTM

Pada gambar 5, Model klasifikasi yang dievaluasi menunjukkan kinerja yang cukup optimal dengan capaian *accuracy* sebesar 90,94%. Pada kelas 0, model menghasilkan nilai *precision* sebesar 0,9598, *recall* sebesar 0,8956, dan *f1-score* sebesar 0,9266, yang mengindikasikan kemampuan tinggi dalam mengidentifikasi kelas ini secara tepat. Sementara itu, untuk kelas 1, *precision* sebesar 0,8351 dan *recall* sebesar 0,9337 menghasilkan *f1-score* sebesar 0,8816, menunjukkan sensitivitas model yang baik meskipun dengan tingkat *precision* yang sedikit lebih rendah. Nilai rata-rata makro untuk *precision*, *recall*, dan *f1-score* masing-masing sebesar 0,8974, 0,9147, dan 0,9041, sedangkan nilai rata-rata tertimbangnyanya adalah 0,9147, 0,9094, dan 0,9103. Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa model mampu melakukan klasifikasi secara efektif dan seimbang, meskipun terdapat ketidak seimbangan jumlah data antar kelas.



Gambar 6. Training/Validation Accuracy dan Loss.

Gambar 6, menunjukkan grafik *accuracy* dan *loss* selama proses pelatihan model IndoBERT-CNN -BiLSTM dalam tugas klasifikasi *gender* berbasis nama. Grafik kiri memperlihatkan peningkatan akurasi baik pada data pelatihan maupun validasi, yang menunjukkan bahwa model belajar secara efektif seiring bertambahnya *epoch*. Akurasi validasi relatif stabil pada kisaran 0,93 hingga 0,94 setelah *epoch* ke-2, menunjukkan performa generalisasi yang baik.

Pada grafik kanan, nilai *loss* pada data pelatihan dan validasi menurun secara konsisten, dengan *loss* validasi yang lebih rendah daripada *loss* pelatihan setelah *epoch* ke-1. Hal ini mengindikasikan bahwa model tidak mengalami *overfitting*, dan justru menunjukkan kemampuan yang baik dalam mengenali pola data secara umum. Secara keseluruhan, kedua grafik memperkuat temuan bahwa arsitektur IndoBERT yang dikombinasikan dengan CNN dan BiLSTM menghasilkan model yang stabil, akurat, dan efisien.

Tabel 7, menunjukkan perbandingan kinerja tiga model berbasis IndoBERT untuk klasifikasi *gender* berdasarkan nama. Model IndoBERT-CNN-BiLSTM memberikan hasil terbaik dengan *accuracy* 90,94%, *precision* 91,47%, *recall* 90,94%, dan *F1-score* 91,03%. Hasil ini menunjukkan bahwa kombinasi IndoBERT, CNN, dan BiLSTM mampu menangkap informasi kontekstual, spasial, dan sekuensial secara lebih baik dibandingkan model lainnya. Dengan performa yang seimbang dan konsisten, model ini dinilai paling optimal untuk tugas klasifikasi *gender* berbasis nama dalam bahasa Indonesia.

Tabel 7. Hasil Perbandingan Model.

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT-CNN	90.52%	91.03%	90.52%	90.26%
IndoBERT-BiLSTM	89.69%	90.56%	89.69%	89.82%
IndoBERT-CNN-BiLSTM	90.94%	91.47%	90.94%	91.03%

3.6. Diskusi

Pemodelan IndoBERT-CNN-BiLSTM dipilih karena mengintegrasikan keunggulan tiga arsitektur yang saling melengkapi dalam klasifikasi teks. IndoBERT berfungsi sebagai ekstraktor fitur kontekstual yang memahami makna kata dalam konteks bahasa Indonesia. CNN digunakan untuk mengekstraksi pola lokal dalam struktur nama, seperti awalan atau akhiran yang berkaitan dengan gender. Sementara itu, BiLSTM digunakan untuk menangkap hubungan sekuensial dua arah dalam rangkaian nama. Kombinasi ini memungkinkan model memproses informasi semantik, spasial, dan temporal secara menyeluruh, sehingga lebih efektif dalam memodelkan karakteristik nama untuk prediksi *gender*.

Pemisahan model IndoBERT-CNN dan IndoBERT-BiLSTM dilakukan untuk mengukur sejauh mana masing-masing komponen berkontribusi terhadap performa model gabungan IndoBERT-CNN-BiLSTM. Dengan membandingkan hasil dari kedua model secara terpisah, dapat diketahui peran CNN dalam mengenali pola lokal dan peran BiLSTM dalam memahami urutan nama. Hasil evaluasi menunjukkan bahwa meskipun model gabungan memberikan hasil terbaik, model parsial juga memiliki performa yang cukup baik, yang menunjukkan bahwa CNN dan BiLSTM saling melengkapi dalam meningkatkan akurasi prediksi *gender*.

Permasalahan umum dalam pembelajaran mendalam, yaitu ketika model terlalu menyesuaikan diri dengan data pelatihan hingga mengurangi kemampuan generalisasi. Meskipun model yang digunakan cukup kompleks karena menggabungkan IndoBERT, CNN, dan BiLSTM, hasil evaluasi menunjukkan tidak terjadinya overfitting. Hal ini ditunjukkan oleh akurasi dan loss yang stabil pada data pelatihan dan validasi, serta nilai loss validasi yang lebih rendah setelah beberapa epoch. Faktor seperti *preprocessing* yang tepat, pembagian data yang seimbang, dan penggunaan IndoBERT yang sesuai dengan bahasa Indonesia turut berperan dalam mencegah overfitting.

Tabel 8. Hasil Perbandingan Dengan Penelitian Terdahulu.

Peneliti	Data	Fokus Penelitian	Hasil
[2]	Dataset UIT-ViNames berisi 26.850 nama lengkap dari mahasiswa di berbagai universitas di Vietnam	Machine Learning: <i>Multinomial Naive Bayes</i> (MNB), <i>Bernoulli Naive Bayes</i> (BNB), <i>Support Vector Machine</i> (SVM), <i>Logistic Regression</i> (LR), <i>Decision Tree</i> (DT), <i>Random Forest</i> (RF). Deep Learning <i>Long Short-Term Memory</i> (LSTM)	Model LSTM dengan <i>fastText</i> menghasilkan F1-score tertinggi sebesar 96% dalam memprediksi <i>gender</i> dari nama-nama Vietnam.
[3]	Dataset berisi 100.787 nama depan dari Brazil berdasarkan sensus 2010 (Brasil.io)	<i>Multilayer Perceptron</i> (MLP), <i>Convolutional Neural Network</i> (CNN), <i>Recurrent Neural Network</i> (RNN), <i>Gated Recurrent Unit</i> (GRU), <i>Bidirectional Long Short-Term Memory</i> (BiLSTM)	Model rekuren seperti RNN dan BiLSTM berhasil memprediksi <i>gender</i> berdasarkan nama dengan akurasi lebih dari 90%, dan mengungguli model <i>feedforward</i> .
[10]	Dataset berisi 5.748 nama lengkap Bengladesh	<i>K-Nearest Neighbor</i> (KNN) <i>Random Forest</i> (RF) <i>Naive Bayes</i> (NB) <i>Decision Tree</i> (DT) <i>Stochastic Gradient Descent</i> (SGD) <i>Logistic Regression</i> (LR) <i>Neural Network</i> (NN)	Model <i>Neural Network</i> dalam penelitian ini mencapai akurasi terbaik sebesar 73,04% untuk klasifikasi <i>gender</i> berdasarkan nama Bengali.
[4]	Dataset Pemilih Pilkada DKI Jakarta 2022 (KPU),	<i>Long Short Term Memory</i> (LSTM)	Penelitian ini menggunakan metode LSTM untuk memprediksi <i>gender</i> dari nama orang Indonesia dan

Peneliti	Data	Fokus Penelitian	Hasil
	serta data mahasiswa Fasilkom Unsri		mencapai akurasi tertinggi sebesar 96% setelah modifikasi model.
Penelitian saat ini	Dataset mahasiswa Universitas Bumigora dari angkatan tahun 2018-2023	Kombinasi Model IndoBERT-CNN, IndoBERT-BiLSTM, dan IndoBERT-CNN-BiLSTM	Model IndoBERT-CNN-BiLSTM berhasil mencapai akurasi 90,94% dan <i>F1-score</i> 91,03% dalam memprediksi <i>gender</i> berdasarkan nama. Model ini stabil dan tidak <i>overfitting</i> selama pelatihan.

4. Kesimpulan

Penelitian ini mengembangkan model klasifikasi jenis kelamin berbasis nama dalam bahasa Indonesia dengan menggabungkan arsitektur IndoBERT, CNN, dan BiLSTM. Model hibrida ini dirancang untuk memanfaatkan kekuatan representasi kontekstual dari IndoBERT, kemampuan ekstraksi fitur lokal dari CNN, serta pemahaman hubungan sekuensial dua arah dari BiLSTM. Berdasarkan hasil evaluasi, model gabungan IndoBERT-CNN-BiLSTM menunjukkan kinerja terbaik dibandingkan dua model pembanding (IndoBERT-CNN dan IndoBERT-BiLSTM), dengan capaian metrik evaluasi yaitu *accuracy* 90,94%, *precision* 91,47%, *recall* 90,94%, dan *F1-score* 91,03%. Stabilitas nilai akurasi dan penurunan nilai loss selama proses pelatihan dan validasi juga mengindikasikan bahwa model tidak mengalami *overfitting* serta mampu melakukan generalisasi dengan baik.

Dengan performa tersebut, model ini dinilai efektif dalam mengatasi permasalahan klasifikasi gender berdasarkan nama yang memiliki struktur tidak terstandarisasi dan dipengaruhi oleh keragaman budaya lokal dalam bahasa Indonesia. Oleh karena itu, pendekatan ini memiliki potensi untuk diterapkan dalam berbagai sistem yang memerlukan prediksi gender berbasis nama, seperti sistem informasi kependudukan, personalisasi layanan, dan analisis demografi. Sebagai rekomendasi untuk pengembangan lebih lanjut, penelitian ini dapat diperluas dengan menggunakan data lintas wilayah dan latar belakang budaya yang lebih beragam guna meningkatkan kemampuan generalisasi model. Selain itu, integrasi model ke dalam sistem berbasis aplikasi atau layanan daring (web-based API) direkomendasikan agar dapat digunakan secara praktis dalam skenario dunia nyata, khususnya dalam pemrosesan data berskala besar secara otomatis dan real-time.

Referensi

- [1] H. N. Trung, V. T. Hoang, and T. H. Huong, "Transformer-based Approach for Gender Prediction using Vietnamese Names," *Procedia Comput. Sci.*, vol. 235, pp. 2362–2369, 2024, <https://doi.org/10.1016/j.procs.2024.04.224>.
- [2] Z. You *et al.*, "Gender Prediction Based on Vietnamese Names with Machine Learning Techniques".
- [3] R. C. B. Rego, V. M. L. Silva, and V. M. Fernandes, "Predicting Gender by First Name Using Character-level Machine Learning," 2021, <http://arxiv.org/abs/2106.10156>
- [4] R. Artikel, A. M. Kusuma, H. Aulia, M. A. Oktavian, M. R. Akbar, and A. Abdiansah, "Prediksi Gender Berdasarkan Nama Bahasa Indonesia Menggunakan Long Short Term Memory Gender Prediction Based on Indonesian Names Using Long Short Term Memory," vol. 9, pp. 265–271, 2023, <https://doi.org/10.28932/jutisi.v9i2.6404>.
- [5] L. Rhue, S. Goethals, and A. Sundararajan, "Evaluating LLMs for Gender Disparities in Notable Persons," 2024, <http://arxiv.org/abs/2403.09148>
- [6] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 1911–1922, 2024, <https://doi.org/10.11591/ijai.v13.i2.pp1913-1924>.
- [7] W. Zichang, L. Xiaoping, D. Science, and C. Science, "Gender prediction model based on CNN-BiLSTM-attention hybrid," vol. 33, no. 4, pp. 2366–2390, <https://www.aimspress.com/aimspress-data/era/2025/4/PDF/era-33-04-105.pdf>.
- [8] E. E. Abdallah, J. R. Alzghoul, and M. Alzghool, "Gender Prediction Based on various Nationality Names using Deep Learning techniques," *Procedia Comput. Sci.*, vol. 170, pp. 563–570, 2020, <https://doi.org/10.1016/j.procs.2020.03.126>.

-
- [9] R. C. B. Rego, V. M. L. Silva, and V. M. Fernandes, "Predicting Gender by First Name Using Character-level Machine Learning," pp. 1–8, 2021, <http://arxiv.org/abs/2106.10156>
 - [10] J. F. Ani, M. Islam, N. J. Ria, S. Akter, and A. K. M. Masum, "Estimating Gender Based On Bengali Conventional Full Name With Various Machine Learning Techniques," *2021 12th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2021*, 2021, <https://doi.org/10.1109/ICCCNT51525.2021.9579927>.
 - [11] R. Ghosh, "Name based Gender Identification using Machine Learning and Deep Learning Models," pp. 1–6, 2023, <https://doi.org/10.36227/techrxiv.21388140.v1>.
 - [12] A. Zein, "Memprediksi Usia Dan Jenis Kelamin Menggunakan Convolutional Neural Networks," *Sainstech J. Penelit. dan Pengkaj. Sains dan Teknol.*, vol. 30, no. 1, pp. 1–7, 2020, <https://doi.org/10.37277/stch.v30i1.727>.
 - [13] A. Saoud, M. Alomeyr, M. F. Amasyali, and H. T. Kesgin, "Inferring gender from first names: Comparing the accuracy of Genderize, Gender API, and the gender R package on authors of diverse nationality," *2024 Innov. Intell. Syst. Appl. Conf. ASYU 2024*, 2024, <https://doi.org/10.1109/ASYU62119.2024.10757039>.
 - [14] R. Ghosh, "Unveiling Gender from Indonesian Names Using Random Forest and Logistic Regression Algorithms," *TechRxiv*, pp. 1–6, 2022.
 - [15] W. Anggraeni, M. F. A. Kusuma, E. Riksakomara, R. P. Wibowo, Pujiadi, and S. Sumpeno, "Combination of BERT and Hybrid CNN-LSTM Models for Indonesia Dengue Tweets Classification," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 1, pp. 813–826, 2024, <https://doi.org/10.22266/ijies2024.0229.68>.
 - [16] H. D. Arpita, A. Al Ryan, M. F. Hossain, M. S. Rahman, M. Sajjad, and N. N. I. Prova, "Exploring Bengali speech for gender classification: machine learning and deep learning approaches," *Bull. Electr. Eng. Informatics*, vol. 14, no. 1, pp. 328–337, 2025, <https://doi.org/10.11591/eei.v14i1.8146>.