KLASIFIKASI LIRIK LAGU MELALUI *SPEECH-TO-TEXT* DENGAN INDOBERT *MULTICLASS CLASSIFICATION* DAN LLAMA3:70B

Michael Natanael
[1]Jurusan Teknik Informatika dan Komputer, Politeknik Negeri Jakarta
Depok, Jawa Barat
michael.natanael.tik21@mhsw.pnj.ac.id

Abstract* - The popularity of Indonesian songs on social media and streaming services has raised concerns regarding age-inappropriate lyrics, particularly their impact on the behavior and emotions of children and adolescents. Therefore, there is a need to develop a model capable of classifying Indonesian song lyrics into four listener age categories: children, adolescents, adults, and all ages. Previous research attempted to classify Indonesian song lyrics by age category using the Naïve Bayes Classifier (NBC) algorithm, yielding unsatisfactory results with an accuracy of 65%. This study proposes a new method that combines a Large Language Model (LLM), specifically LLAMA3:70B, to improve classification performance and expand the dataset by up to 1000%. Three dataset augmentation strategies were applied in this method: zero-shot, sample-based, and translation-based. Classification performance was significantly improved through the best model, a fine-tuned IndoBERT multiclass classification with a sample-based approach, which achieved an F1 macro score of 74%. This model was then implemented as a web application equipped with a speech-to-text feature using Whisper-Large-V3. Test results show that the classification accuracy reached 75% on the benchmark dataset. The usability evaluation resulted in a System Usability Scale (SUS) score of 78.21 and a Net Promoter Score (NPS) of 45%.

Keywords: IndoBERT multiclass classification, Listener age categories, Indonesian song lyrics classification, LLAMA3, Whisper

***Abstrak*--** **Popularitas lagu-lagu Indonesia di media sosial dan layanan *streaming* telah menimbulkan kekhawatiran terkait lirik yang tidak sesuai usia, terutama dampaknya terhadap perilaku dan emosi anak-anak dan remaja. Oleh karena itu, diperlukan pengembangan model yang mampu mengklasifikasikan lirik lagu Indonesia ke dalam empat kategori usia pendengar: anak-anak, remaja, dewasa, dan semua usia. Penelitian terdahulu telah melakukan klasifikasi lirik lagu Indonesia berdasarkan kategori usia menggunakan algoritma Naïve Bayes *Classifier* (NBC), menghasilkan hasil yang kurang memuaskan dengan akurasi 65%. Penelitian ini mengusulkan metode baru yang menggabungkan *Large Language Model* (LLM), khususnya LLAMA3:70B untuk meningkatkan kinerja klasifikasi dan memperluas jumlah *dataset* hingga 1000%. Tiga strategi augmentasi *dataset* diterapkan dalam metode ini, yaitu *zero-shot, sample-based,* dan *translation-based.* Performa klasifikasi berhasil ditingkatkan secara signifikan melalui model terbaik hasil *fine-tune* IndoBERT *multiclass classification* dengan pendekatan *sample-based*, yang mencapai *F1 macro score* sebesar 74%. Model ini kemudian diimplementasikan dalam bentuk aplikasi web yang dilengkapi dengan fitur *speech-to-text* dengan Whisper-Large-V3. Hasil pengujian menunjukkan bahwa akurasi klasifikasi mencapai 75% pada *benchmark dataset*. Hasil evaluasi *usability* menunjukkan nilai *System Usability Scale* (SUS) sebesar 78,21 dan *Net Promoter Score* (NPS) sebesar 45%.**

Kata kunci: IndoBERT multiclass classification, Kategori usia pendengar, Klasifikasi lirik lagu Indonesia, LLAMA3, Whisper

I. PENDAHULUAN

Perkembangan teknologi memberikan pengaruh besar pada berbagai industri kreatif, khususnya pada industri musik Indonesia karena tingginya penetrasi media sosial dan layanan *streaming* [1]. Survei menunjukkan bahwa mayoritas masyarakat Indonesia secara rutin mendengarkan musik melalui layanan *streaming* [2] [3]. Namun, kemudahan akses terhadap jutaan lagu ini tidak diimbangi dengan sistem klasifikasi konten yang memadai. Akibatnya, anak-anak dan remaja dapat dengan mudah mengakses lagu dengan lirik

yang mengandung bahasa kasar, diskriminatif, atau referensi kekerasan dan seksual tanpa pengawasan orang dewasa [4]. Hal ini sangat kontras dengan industri film yang telah menerapkan label klasifikasi usia secara jelas dalam memilih tontonan yang sesuai [5]. Paparan lirik bertema dewasa ini dapat meningkatkan kecenderungan perilaku berbahaya, seperti agresi atau kekerasan, bunuh diri, penggunaan tembakau, alkohol, dan obat-obatan terlarang, serta praktik seksual yang tidak aman [6].

Saat ini, telah ada penelitian sejenis yang mengembangkan klasifikasi lirik lagu Indonesia menggunakan metode Naive Bayes *Classifier*. Namun, penelitian tersebut hanya mencapai akurasi sebesar 65% dengan rasio 80% data latih dan 20% data uji [5]. Rendahnya akurasi ini disebabkan oleh keterbatasan metode Naive Bayes *Classifier* yang mengasumsikan kelompok usia lirik dengan menganalisis frekuensi kata dan memperlakukan setiap kata sebagai fitur independen sehingga gagal memahami hubungan semantik dan kontekstual dalam lirik lagu, serta kurangnya jumlah *dataset* yang tersedia untuk pelatihan model [7].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan sistem klasifikasi lirik lagu Indonesia berdasarkan kelompok usia dengan mengintegrasikan teknologi *speech-to-text* dan IndoBERT *Pre-trained Language Model* (PLM). IndoBERT dipilih karena kemampuannya dalam memahami hubungan kontekstual antar kata melalui mekanisme *self-attention* [8]. LLAMA3:70B digunakan untuk menghasilkan *dataset* augmentasi lirik yang mengadopsi mekanisme generasi lirik dan metode pengambilan sampel [9] [10]. Kemudian IndoBERT *multiclass classification* digunakan untuk mengklasifikasikan lirik lagu berdasarkan kelompok usia, meliputi anak-anak, remaja, dewasa, dan semua usia. Setelah itu, model yang telah dikembangkan diimplementasikan ke dalam bentuk aplikasi web yang terdapat fitur *speech-to-text* untuk mengubah audio menjadi lirik lagu.

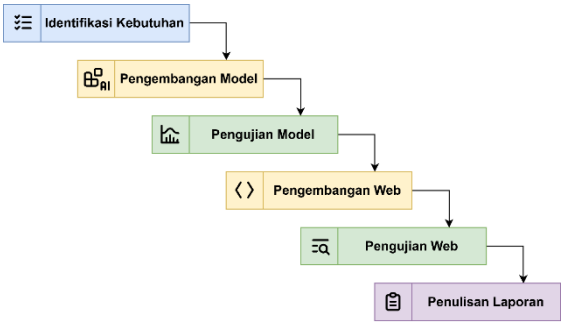
II. METODOLOGI PENELITIAN

A. Rancangan Penelitian

Penelitian ini bertujuan untuk mengembangkan dan mengimplementasikan sistem klasifikasi lirik lagu Indonesia berdasarkan kelompok usia. Sistem ini mengintegrasikan teknologi *speech-to-text* menggunakan Whisper, model klasifikasi IndoBERT *multiclass classification*, serta LLAMA3:70B untuk augmentasi *dataset*. Tujuan utama penelitian ini adalah mengembangkan sistem yang mampu secara otomatis mengklasifikasikan lirik lagu Indonesia berdasarkan kategori usia pendengar dengan tingkat akurasi yang tinggi. Sistem ini diharapkan dapat membantu melakukan klasifikasi lirik lagu dari teks atau audio yang diunggah ke dalam aplikasi berbasis web.

B. Tahapan Penelitian

Tahapan penelitian yang telah diterapkan diilustrasikan pada Gambar 1.



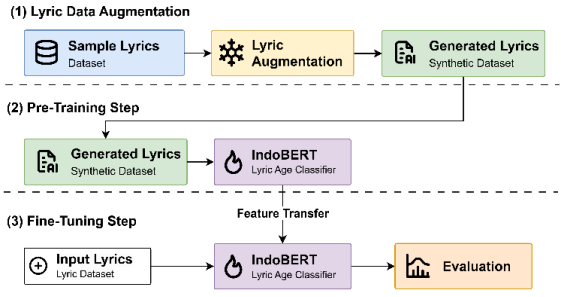
Gambar 1. Tahapan Penelitian

1. Identifikasi Kebutuhan

Penelitian diawali dengan pengumpulan dataset lirik lagu Indonesia dan perumusan masalah mengenai akses anak-anak dan remaja terhadap konten yang tidak sesuai usia. Selanjutnya, dilakukan studi literatur terkait teknologi Whisper, IndoBERT, dan LLAMA3:70B untuk menetapkan ruang lingkup dan batasan penelitian.

2. Pengembangan Model

Data lirik lagu yang telah dikumpulkan akan melalui tahap pembersihan dan pra-pemrosesan agar dapat digunakan dalam pelatihan model. Pada tahap ini, LLAMA3:70B digunakan untuk melakukan augmentasi *dataset* dan IndoBERT *multiclass classification* untuk melatih model klasifikasi lirik lagu berdasarkan kelompok usia. sebagaimana ditampilkan pada Gambar 2.



Gambar 2. Tahapan Metode Klasifikasi Lirik Berdasarkan Usia

Metode klasifikasi lirik berdasarkan kelompok usia terdiri atas tiga langkah utama, yaitu

lyric data augmentation, pre-training step, dan *fine-tuning step*.

3. Pengujian Model

Tahap berikutnya adalah pengujian model yang telah dikembangkan. Kinerja model dievaluasi menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan Kappa untuk memastikan akurasi dan efektivitas sistem dalam mengklasifikasikan lirik lagu.

4. Pengembangan Web

Setelah model selesai diuji, penelitian berlanjut ke tahap pengembangan aplikasi web. Sistem klasifikasi yang telah dibangun diimplementasikan ke dalam aplikasi web dengan menambahkan fitur *speech-to-text* menggunakan Whisper.

5. Pengujian Web

Tahap pengujian web dilakukan untuk memastikan bahwa aplikasi web berjalan dengan baik dan fitur-fitur yang disediakan berfungsi sebagaimana mestinya, seperti *speech-to-text* dan klasifikasi lirik lagu berdasarkan kelompok usia. Pengujian ini melibatkan metode *black box testing*, *benchmark dataset evaluation*, *System Usability Scale* (SUS), dan *Net Promoter Score* (NPS).

6. Penulisan Laporan

Penelitian ini ditutup dengan penyusunan laporan akhir yang mendokumentasikan seluruh proses penelitian, mulai dari identifikasi kebutuhan, pengembangan dan pengujian model, hingga implementasi dan pengujian aplikasi web. Laporan ini juga mencakup analisis hasil, tantangan yang dihadapi, dan rekomendasi untuk pengembangan lebih lanjut di masa depan.

C. Objek Penelitian

Objek penelitian ini adalah *dataset* berisi 400 lirik lagu Indonesia yang telah dikelompokkan berdasarkan usia pendengar, yaitu anak-anak, remaja, dewasa, dan semua usia. Masing-masing kategori terdiri dari 100 lirik lagu. *Dataset* ini diperoleh dari penelitian terdahulu [5].

II. HASIL DAN PEMBAHASAN

A. Analisis Kebutuhan

Analisis kebutuhan dilakukan untuk menentukan fitur dan teknologi yang diperlukan dalam pengembangan sistem klasifikasi lirik lagu berbasis web. Kebutuhan fungsional sistem dapat dilihat pada Tabel 1.

Tabel 1. Kebutuhan Fungsional

Nama	Deskripsi
Input Lirik	Pengguna dapat mengunggah *file* audio (.mp3) atau memasukkan teks lirik manual.
Speech-to-Text (STT)	Konversi otomatis audio ke teks menggunakan API Whisper (OpenAI).
Prediksi Usia	Menampilkan hasil klasifikasi usia dengan probabilitas prediksi.
Riwayat Prediksi	Penyimpanan riwayat prediksi beserta metadata (lirik, tanggal, durasi proses, dll).
Ekspor Hasil	Ekspor hasil ke format CSV, JSON, TXT, atau SQL untuk analisis lebih lanjut.

Selain kebutuhan fungsional, sistem juga harus memenuhi beberapa kebutuhan non-fungsional berikut untuk meningkatkan kenyamanan penggunaan pada Tabel 2.

Tabel 2. Kebutuhan Non-Fungsional

Nama	Deskripsi
Responsivitas	Antarmuka web responsif di perangkat *mobile* dan *desktop*.
Kecepatan Prediksi	Proses prediksi selesai dalam ≤ 10 detik (tidak termasuk fitur *speech-to-text*).
Keamanan Data	Enkripsi data pengguna terdaftar (SHA256).

Teknologi yang akan digunakan untuk membangun sistem ini meliputi:

1. Pengembangan Model

Menggunakan library Python seperti PyTorch, PyTorch Lightning, HuggingFace Transformers, Pandas, NumPy, dan Groq untuk mengakses API LLM.

2. Pengembangan Web

Menggunakan *framework* Flask, teknologi *containerization* Docker, dan *source-code editor* VSCode.

B. Perancangan Sistem

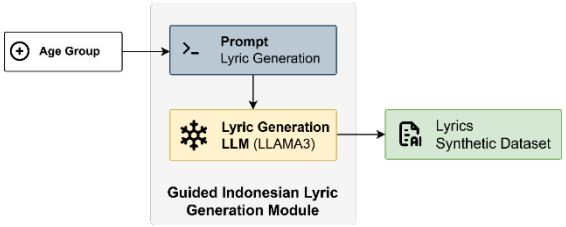
Perancangan sistem dibagi menjadi dua bagian utama: perancangan model klasifikasi dan perancangan aplikasi web.

1. Perancangan Model

Dataset awal dari penelitian sebelumnya sebanyak 400 lirik, dilakukan augmentasi dengan LLAMA3:70B untuk mencapai target 1.000 lirik per

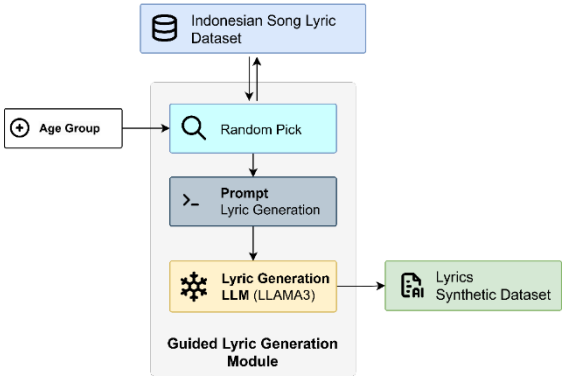
kategori. Terdapat tiga strategi augmentasi dengan LLAMA3:70B sebagai berikut.

Zero-Shot: LLM menghasilkan lirik hanya berdasarkan input kelompok usia tanpa contoh tambahan yang ditunjukkan pada Gambar 3.



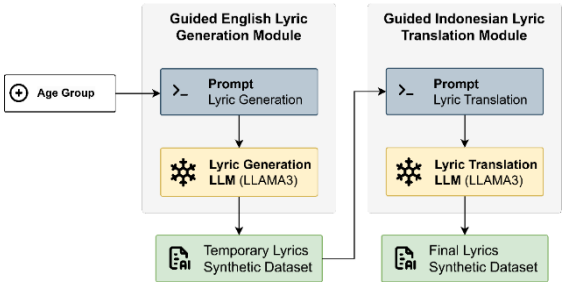
Gambar 3. *Zero-Shot Lyric Augmentation*

Sample-Based: LLM menghasilkan lirik baru dengan diberikan 20 sampel acak dari dataset asli sebagai referensi yang ditunjukkan pada Gambar 4.



Gambar 4. *Sample-Based Lyric Augmentation*

Translation-Based: LLM menghasilkan lirik dalam Bahasa Inggris terlebih dahulu, kemudian diterjemahkan ke Bahasa Indonesia untuk menjaga kreativitas seperti yang ditunjukkan pada Gambar 5.

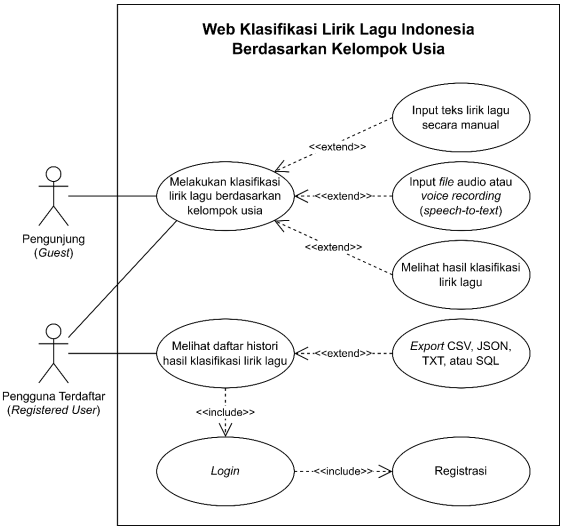


Gambar 5. *Translation-Based Lyric Augmentation*

Kemudian, kualitas lirik hasil augmentasi diuji kelayakannya menggunakan BERTScore untuk menghitung kesamaan semantik antara lirik sintetis dengan lirik asli. Model IndoBERT dilatih dalam dua tahap. Pada tahap *pre-training*, model dilatih pada dataset sintetis hasil augmentasi untuk mempelajari karakteristik lirik spesifik usia. Pada tahap *fine-tuning*, model akan mentransfer fitur-fitur yang telah dilakukan *pre-training* pada *dataset* lirik sintetis kemudian disempurnakan menggunakan *human-annotated lyrics dataset* sebagai *benchmark* untuk meningkatkan akurasi klasifikasi. Dalam penelitian ini, model Whisper-Large-V3 digunakan karena memberikan akurasi transkripsi terbaik untuk Bahasa Indonesia.

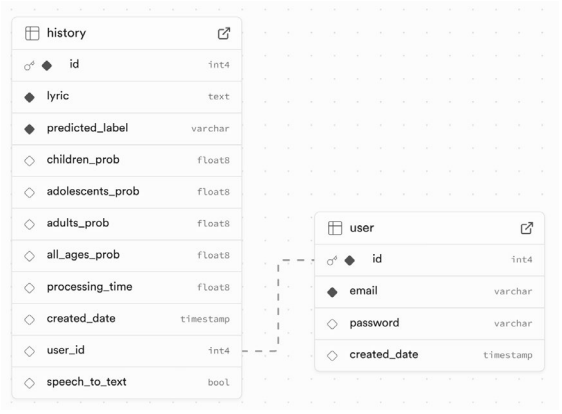
2. Perancangan Web

Gambar 6 menunjukkan *use case diagram* dari sistem web klasifikasi lirik lagu Indonesia berdasarkan kelompok usia yang dikembangkan dalam penelitian ini.



Gambar 6. *Use Case Diagram*

Use case diagram ini dapat menggambarkan hubungan antara aktor dengan aksi yang dapat dilakukan pada web. Terdapat dua jenis aktor utama dalam sistem, yaitu Pengunjung (*Guest*) dan Pengguna Terdaftar (*Registered User*).



Gambar 7. Struktur *Database* PostgreSQL

Database yang digunakan dalam penelitian ini adalah PostgreSQL yang dihosting melalui layanan Supabase. Struktur *database* pada penelitian ini ditampilkan pada Gambar 7, yang terdiri dari dua tabel, yaitu tabel *user* dan *history*. Tabel *user* berfungsi untuk menyimpan data akun pengguna. Tabel *history* menyimpan riwayat klasifikasi lirik lagu yang dilakukan oleh pengguna.

C. Implementasi Sistem

Berbagai rancangan yang telah dibuat kemudian direalisasikan dalam tahap implementasi sistem. Dalam penelitian ini, implementasi sistem dibagi menjadi dua tahap, yaitu implementasi model dan implementasi web.

1. Implementasi Model

LLAMA 3.70B digunakan pada ketiga strategi augmentasi *dataset* lirik Indonesia, yaitu *zero-shot*, *sample-based*, dan *translation-based*. Jumlah lirik sintetis hasil augmentasi *dataset* dapat dilihat pada Tabel 3.

Tabel 3. Jumlah Lirik Sintetis Hasil Augmentasi *Dataset*

Strategi Augmentasi	**Jumlah Lirik Sintetis Berdasarkan Kategori Usia**				**Total**
	Anak	**Remaja**	**Dewasa**	**Semua Usia**	
Zero-Shot	1003	1001	1003	1002	4009
Sample-Based	1002	1003	1001	1000	4006
Translation-Based	1000	1000	1000	1003	4003

Dataset sintetis hasil augmentasi selanjutnya dilakukan uji kelayakan dataset dengan cara menguji *similarity score*. Hasil Uji *Similarity Score* dengan BERTScore dapat dilihat pada Tabel 4.

Tabel 4. Hasil Uji *Similarity Score* dengan BERTScore

***Dataset* Sintetis**	***Precision***	***Recall***	***F1-Score***
Zero-Shot	0,66	0,66	0,66
Sample-Based	0,67	0,66	0,66
Translation-Based	0,67	0,67	0,67

Walaupun pendekatan *translation-based* memberikan hasil *similarity score* terbaik dibandingkan dua metode lainnya, hal ini belum tentu menjamin performa klasifikasi terbaik. Oleh karena itu, analisis selanjutnya akan difokuskan pada performa klasifikasi dengan IndoBERT.

Model IndoBERT dilatih menggunakan pendekatan "*multiclass*" karena tujuan utama dari penelitian ini adalah mengklasifikasikan lirik lagu ke dalam satu dari beberapa kategori usia yang saling eksklusif, yaitu anak-anak, remaja, dewasa, dan semua usia. Selama proses pelatihan juga diterapkan teknik *early stopping*, yaitu teknik untuk menghentikan proses pelatihan apabila performa model sudah tidak bertambah baik. Selain itu, *hyperparameter* merupakan bagian yang wajib ditentukan dalam melatih model IndoBERT. Pengaturan *hyperparameter* ini berdampak langsung pada performa, stabilitas, dan efisiensi konvergensi model selama proses pelatihan.

Tabel 5. *Hyperparameter* Pelatihan Model

Hyperparameter	**Nilai**
Max Length	100
Batch Size	90
Learning Rate	1e-5

Rincian *hyperparameter* untuk *pre-training* dan *fine-tuning* IndoBERT dapat dilihat pada Tabel 5.

Tabel 6. Hasil *Pre-training* IndoBERT

Training Dataset	***Approach***	***Pre-training* IndoBERT (*Save Checkpoint*)**				
		Accuracy	***F1 Macro***	***Prec Macro***	***Recall Macro***	**Kappa**
Generate Indonesian Lyrics	*Zero-Shot*	0.43	0.17	0.26	0.19	0.05
Example + Generate Indonesian Lyrics	*Sample-Based*	**0.54**	**0.27**	**0.34**	**0.28**	**0.10**
Generate English Lyrics + Translation	*Translation-Based*	0.50	0.19	0.25	0.21	0.05

Seperti yang ditunjukkan pada Tabel 6, pendekatan augmentasi berbasis sampel mencapai kinerja tertinggi pada langkah pra-pelatihan. Hal ini terjadi karena memberikan contoh lirik sebagai masukan membantu mengurangi masalah halusinasi dengan menawarkan dasar kontekstual, mengurangi kemungkinan halusinasi, dan meningkatkan kualitas lirik yang dihasilkan.

Tabel 7. Hasil *Fine-tuning* IndoBERT

Methods	***Model***	***Training Dataset***	**GD?**	***Fine-tuning* IndoBERT (*Load From Checkpoint*)**				
				Acc	***F1 Macro***	***Prec Macro***	***Rec Macro***	**Kappa**
Baseline	Naïve Bayes	*Indonesian Lyrics*	✗	0.65	0.64	0.67	0.63	0.53
PLM	Indo BERT	*Indonesian Lyrics*	✗	0.71	0.71	0.72	0.74	0.61
Zero-Shot	Indo BERT	*Generate Indonesian Lyrics*	✓	0.56	0.56	0.61	0.61	0.42
Sample-Based	Indo BERT	*Example + Generate Indonesian Lyrics*	✓	**0.75**	**0.74**	**0.74**	**0.75**	**0.66**
Translation-Based	Indo BERT	*Generate English Lyrics + Translation*	✓	0.69	0.70	0.70	0.71	0.59

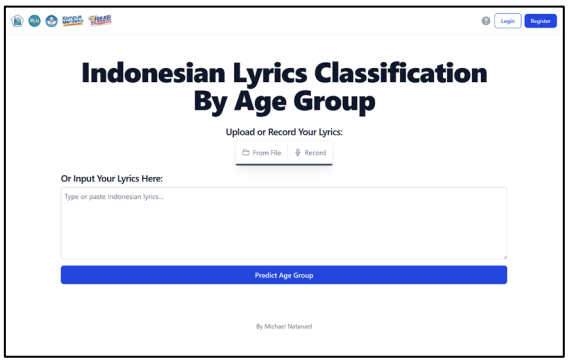
Hasil *fine-tuning* pada Tabel 7 menunjukkan bahwa model IndoBERT mengungguli model Naïve Bayes dalam hal penggunaan *human-annotated dataset* dan *Generated Dataset* (GD). Model IndoBERT dengan pendekatan *Sample-Based* menunjukkan performa terbaik secara keseluruhan dengan akurasi 75% dan F1-Score 74%, mengungguli model *baseline* Naïve Bayes (akurasi 65%). Model inilah yang dipilih untuk dilanjutkan ke tahap implementasi web. Fitur *speech-to-text* diimplementasikan menggunakan model Whisper-Large-V3 melalui API Groq dengan pengaturan parameter lengkap pada Tabel 8.

Tabel 8. *Hyperparameter* Whisper

Hyperparameter	**Nilai**	**Deskripsi**
model	"whisper-large-v3"	Model Whisper yang digunakan untuk transkripsi, mendukung multibahasa.
prompt	"Transkripsikan hanya bagian lirik lagu saja"	Instruksi khusus untuk memandu model fokus pada bagian lirik lagu.
language	"id"	Bahasa audio yang akan ditranskripsi, dalam hal ini Bahasa Indonesia.
temperature	0	Menentukan tingkat *randomness*.

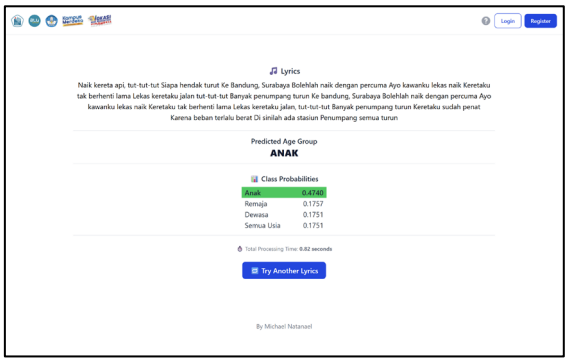
2. Implementasi Web

Halaman utama pada Gambar 8 menampilkan antarmuka yang sederhana untuk input lirik melalui teks, unggahan file, atau rekaman suara.



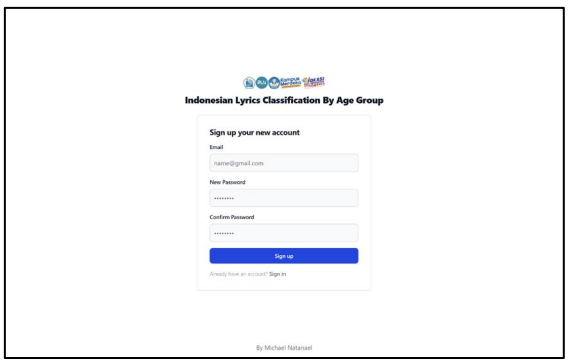
Gambar 8. Halaman Utama Klasifikasi Lirik Lagu

Halaman Hasil Prediksi pada Gambar 9 menampilkan lirik, hasil prediksi utama (misalnya, ANAK), tabel probabilitas untuk semua kelas, dan total waktu pemrosesan.



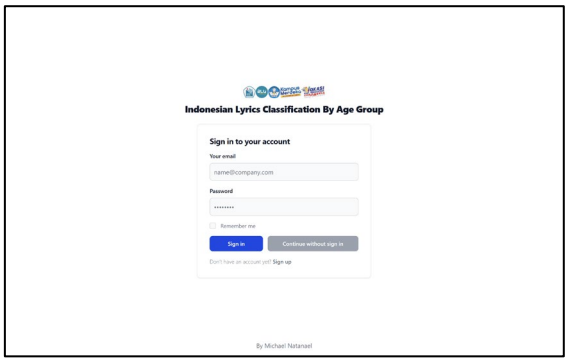
Gambar 9. Halaman Hasil Prediksi Usia

Halaman Registrasi pada Gambar 10 menyediakan formulir standar untuk membuat akun baru.



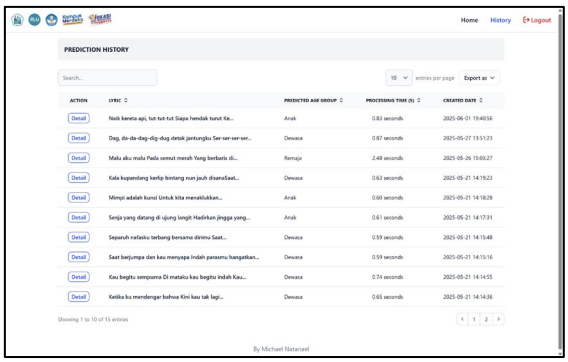
Gambar 10. Halaman Registrasi Akun

Pengguna yang sudah memiliki akun dapat masuk ke dalam sistem melalui halaman *login* seperti yang ditampilkan pada Gambar 11.



Gambar 11. Halaman *Login* Pengguna

Halaman Riwayat Prediksi pada Gambar 12 hanya dapat diakses pengguna terdaftar, menampilkan tabel riwayat yang interaktif dengan fitur pencarian, detail, dan ekspor data.



Gambar 12. Halaman Riwayat Prediksi

D. Pengujian Sistem

Pengujian dilakukan untuk mengevaluasi kualitas sistem secara fungsional, performa model, dan kebergunaan dari sisi pengguna.

1. Deskripsi Pengujian

Pengujian terdiri dari empat metode sebagai berikut.

Black Box Testing: Untuk verifikasi fungsionalitas sistem.

Benchmark Dataset Evaluation: Untuk mengukur performa model IndoBERT pada data anotasi manual.

System Usability Scale (SUS): Untuk mengukur tingkat kebergunaan sistem.

Net Promoter Score (NPS): Untuk mengukur kesediaan pengguna merekomendasikan sistem.

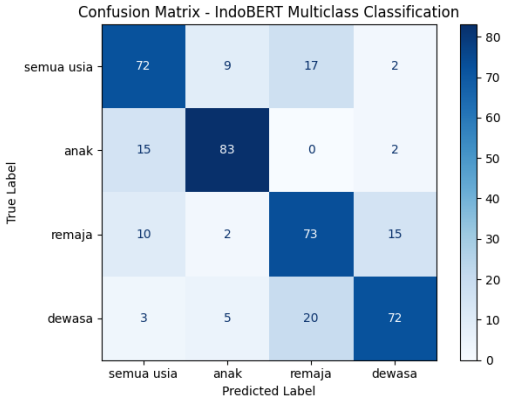
2. Prosedur Pengujian

Black box dan *benchmark evaluation* dilakukan secara mandiri. Sementara itu, SUS dan NPS diuji melalui kuesioner yang disebarkan kepada 42 responden dari berbagai latar belakang usia, jenis kelamin, dan pekerjaan.

3. Data Hasil Pengujian

Black Box Testing: Seluruh skenario pengujian fungsionalitas berhasil 100% sesuai dengan hasil yang diharapkan.

Benchmark Dataset Evaluation: Con*fusion matrix* pada Gambar 13 menunjukkan model memiliki akurasi keseluruhan 75%. Kategori "anak" memiliki prediksi benar tertinggi (83 kasus), sedangkan kesalahan klasifikasi paling sering terjadi antara kategori "remaja" dan "dewasa".



Gambar 13. *Confusion Matrix* IndoBERT *Multiclass Classification*

System Usability Scale (SUS): Rata-rata skor SUS dari 42 responden adalah 78,21, yang masuk dalam kategori "*Good*" dengan *grade* "B".

Net Promoter Score (NPS): Nilai NPS yang didapat adalah 45%, yang tergolong dalam kategori "*Great*" dan menunjukkan bahwa pengguna puas serta bersedia merekomendasikan sistem.

4. Analisis Data Pengujian

Analisis Kesalahan Prediksi dengan *Explainable* AI: Menggunakan LIME dan SHAP, diidentifikasi bahwa kesalahan prediksi seringkali disebabkan oleh keterbatasan dataset pelatihan. Misalnya, lagu "Satu Satu Aku Sayang Ibu" salah diklasifikasikan sebagai "dewasa" karena kata "sayang" lebih sering muncul dalam konteks romantis di data pelatihan.

Analisis Demografi SUS: Responden dewasa (86,30) dan berprofesi sebagai Guru/Dosen

(100) memberikan skor SUS tertinggi, sementara kelompok usia anak-anak (67,50) dan remaja (69,25) memberikan skor lebih rendah. Ini mengindikasikan antarmuka lebih mudah dipahami oleh pengguna dewasa.

Analisis Demografi NPS: Kelompok anak-anak (67%) dan profesi Guru/Dosen serta Wiraswasta (100%) menunjukkan kesediaan tertinggi untuk merekomendasikan sistem. Sebaliknya, kelompok remaja (20%) dan Karyawan (-25%) lebih kritis atau kurang puas.

III. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem klasifikasi lirik lagu Indonesia berdasarkan usia menggunakan model IndoBERT yang mencapai F1 Macro 74% dan akurasi 75% pada pengujian. Sistem ini diimplementasikan dalam sebuah aplikasi web fungsional yang dilengkapi fitur *speech-to-text* menggunakan Whisper-Large-V3, riwayat prediksi, dan autentikasi pengguna. Meskipun pengujian pengalaman pengguna menunjukkan hasil positif dengan skor SUS 78,21 dan NPS 45%, analisis XAI (LIME dan SHAP) mengungkap bahwa kesalahan klasifikasi utamanya disebabkan oleh keterbatasan representasi data latih pada lirik bertema kompleks atau eksplisit. Oleh karena itu, untuk pengembangan selanjutnya, disarankan untuk meningkatkan kualitas model melalui pendekatan *Reinforcement Learning from Human Feedback* (RLHF) yang melibatkan pakar, memperkaya fitur dengan menambahkan fungsi identifikasi lagu, serta melakukan pengujian pada populasi yang lebih luas untuk meningkatkan generalisasi dan *fairness* model. Implementasi saran-saran ini diharapkan dapat meningkatkan kualitas dan fungsionalitas sistem agar dapat memberikan manfaat yang lebih besar bagi dunia edukasi dan industri musik digital di Indonesia.

IV. REFERENSI

[1] S. Minarni dan K. S. Dewi, "PENGARUH WORD of MOUTH, DIGITAL MARKETING, BRAND AWARENESS TERHADAP COSTUMER LOYALTY PADA ARTIS BARU SEVEN MUSIC INDONESIA," *JMB*, vol. 3, no. 1, hlm. 1–17, Apr 2023, doi: 10.32509/jmb.v3i1.2683.

[2] A. Fauzan, R. Cleverin, dan S. B. P. Lumbaa, "THE EFFECTS OF DIGITAL PLATFORMS IN THE MUSIC INDUSTRY TO R&B MUSIC IN INDONESIA," vol. 12, no. 2, 2023.

[3] IDN Research Institute, "INDONESIA GEN Z REPORT 2024 (Understanding and Uncovering the Behavior, Challenges, and Opportunities)," IDN Media, 2024. Diakses: 9 Januari 2025. [Daring]. Tersedia pada: https://cdn.idntimes.com/content-documents/indonesia-gen-z-report-2024.pdf

[4] M. Rospocher dan S. Eksir, "Assessing Fine-Grained Explicitness of Song Lyrics," *Information*, vol. 14, no. 3, hlm. 159, Mar 2023, doi: 10.3390/info14030159.

[5] Moch. F. Shadiqin Thirafi dan F. Rahutomo, "Implementation of Naïve Bayes Classifier Algorithm to Categorize Indonesian Song Lyrics Based on Age," dalam *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia: IEEE, Nov 2018, hlm. 106–109. doi: 10.1109/SIET.2018.8693201.

[6] E. Kury, E. Kury, N. Quinn, dan R. P. Olympia, "Lyrical Content of Contemporary Popular Music (1999-2018) and the Role of Healthcare Providers in Media Education of Children and Adolescents," *Cureus*, Okt 2020, doi: 10.7759/cureus.11214.

[7] A. Kelly dan M. A. Johnson, "Investigating the Statistical Assumptions of Naïve Bayes Classifiers," dalam *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA: IEEE, Mar 2021, hlm. 1–6. doi: 10.1109/CISS50987.2021.9400215.

[8] F. Koto, A. Rahimi, J. H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," dalam *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, hlm. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[9] N. Jonason, L. Casini, C. Thomé, dan B. L. T. Sturm, "Retrieval Augmented Generation of Symbolic Music with LLMs," 28 Desember 2023, *arXiv*: arXiv:2311.10384. Diakses: 18 Agustus 2024. [Daring]. Tersedia pada: http://arxiv.org/abs/2311.10384

[10] S. Liu, A. S. Hussain, C. Sun, dan Y. Shan, "M^{2}UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models," 4 Maret 2024, *arXiv*: arXiv:2311.11255. Diakses: 21 Agustus 2024. [Daring]. Tersedia pada: http://arxiv.org/abs/2311.11255