

Model Rekomendasi Musik Berbasis Representasi Semantik Lirik Lagu Menggunakan BERT

¹Dziaul Hululiah, ²Fersellia

^{1,2}Program Studi Teknik Informatika, Fakultas Teknik, Universitas Ma'arif Nahdlatul Ulama Kebumen, Indonesia

¹ziaaava@gmail.com; ²fersellia98@gmail.com;

Article Info

Article history:

Received, 2026-01-14

Revised, 2026-01-27

Accepted, 2026-01-30

Kata Kunci:

rekomendasi music
analisis lirik
BERT
NLP

Keywords:

music recommendation
lyric analysis
BERT
NLP

ABSTRAK

Pertumbuhan pesat *platform* musik digital menimbulkan permasalahan *information overload* yang menyulitkan pengguna dalam menemukan lagu sesuai dengan preferensi mereka. Penelitian ini mengusulkan model rekomendasi musik berbasis konten melalui analisis semantik lirik lagu dengan pendekatan *Natural Language Processing* menggunakan model *Bidirectional Encoder Representations from Transformers*. Tahapan penelitian meliputi pengumpulan data lirik lagu berbahasa Indonesia, pembersihan data, prapemrosesan teks, pembentukan representasi embedding lirik secara kontekstual, serta perhitungan kemiripan antar lirik menggunakan *cosine similarity*. Kinerja model dievaluasi menggunakan metrik *Mean Squared Error* dan akurasi. Hasil eksperimen menunjukkan bahwa model yang diusulkan mencapai tingkat akurasi sebesar 83,69% dengan nilai *Mean Squared Error* sebesar 1,4066. Hasil ini mengindikasikan bahwa representasi lirik berbasis *Bidirectional Encoder Representations from Transformers* mampu menangkap makna semantik lirik secara efektif dan meningkatkan relevansi rekomendasi musik secara kuantitatif. Dengan demikian, pendekatan analisis semantik berbasis lirik berpotensi meningkatkan akurasi dan personalisasi sistem rekomendasi musik berbasis konten.

ABSTRACT

The rapid growth of digital music platforms has resulted in an information overload problem, making it difficult for users to discover songs that match their preferences. This study proposes a content-based music recommendation model through semantic analysis of song lyrics using a Natural Language Processing approach with Bidirectional Encoder Representations from Transformers. The research stages include Indonesian song lyric data collection, data cleaning, text preprocessing, contextual lyric embedding generation, and lyric similarity computation using cosine similarity. Model performance is evaluated using Mean Squared Error and accuracy. Experimental results show that the proposed model achieves an accuracy of 83.69% with a Mean Squared Error value of 1.4066, indicating that lyric representations generated by Bidirectional Encoder Representations from Transformers effectively capture semantic meaning and quantitatively improve the relevance of music recommendations. Therefore, the proposed approach enhances the accuracy and personalization of content-based music recommendation systems.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-nc-nd/4.0/) license.



Penulis Korespondensi:

Dziaul Hululiah,
Program Studi Teknik Informatika,
Universitas Ma'arif Nahdlatul Ulama Kebumen,
Email: ziaaava@gmail.com

1. PENDAHULUAN

Perkembangan *platform* musik digital seperti *Spotify*, *YouTube Music*, dan *Apple Music* menyebabkan permasalahan *information overload*, di mana pengguna kesulitan menemukan lagu yang sesuai dengan preferensi mereka akibat banyaknya pilihan yang tersedia [1]. Oleh karena itu, sistem rekomendasi menjadi komponen penting untuk membantu pengguna memperoleh saran musik yang relevan dan personal.

Pendekatan *content-based filtering* banyak digunakan dalam sistem rekomendasi musik karena mampu merekomendasikan lagu berdasarkan kemiripan konten, termasuk melalui analisis lirik [2]. Lirik lagu memiliki kekayaan makna semantik dan konteks linguistik yang dapat merepresentasikan karakteristik lagu secara lebih mendalam dibandingkan fitur audio atau metadata sederhana [3]. Penelitian sebelumnya umumnya menggunakan metode representasi teks tradisional seperti *TF-IDF* dan *Word2Vec* untuk menganalisis lirik lagu. Namun, metode tersebut masih memiliki keterbatasan dalam menangkap makna kontekstual secara dua arah, khususnya pada teks lirik yang bersifat kompleks dan puitis [4]. Model berbasis *transformer* seperti *Bidirectional Encoder Representations from Transformers (BERT)* menawarkan kemampuan representasi semantik yang lebih kontekstual, tetapi pemanfaatannya dalam sistem rekomendasi musik berbasis lirik berbahasa Indonesia masih terbatas. Berdasarkan hal tersebut, *research gap* penelitian ini terletak pada belum optimalnya pemanfaatan representasi semantik kontekstual berbasis *BERT* dalam sistem rekomendasi musik berbasis lirik. *Novelty* penelitian ini adalah penerapan *embedding* lirik berbasis *BERT* pada sistem rekomendasi musik berbasis *content-based filtering* untuk lirik lagu berbahasa Indonesia, serta evaluasinya menggunakan metrik *Mean Squared Error*, *precision*, dan *accuracy*. Pendekatan ini diharapkan mampu meningkatkan relevansi dan personalisasi rekomendasi musik [5].

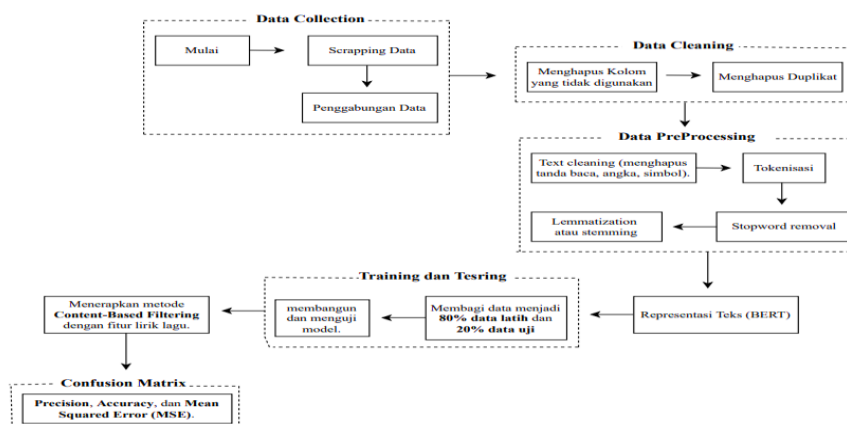
2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Research and Development (R&D)* yang bertujuan untuk mengembangkan dan mengevaluasi model rekomendasi musik berbasis analisis lirik lagu. Pendekatan ini dipilih karena memungkinkan pengembangan model secara sistematis melalui tahapan perancangan, implementasi, serta evaluasi kinerja berdasarkan metrik kuantitatif [6].

Alur penelitian meliputi pengumpulan data, pembersihan data, prapemrosesan teks, pembentukan *embedding* lirik, pelatihan dan pengujian model, perhitungan kemiripan semantik, serta evaluasi kinerja sistem. Alur proses penelitian ditunjukkan pada Gambar 1, sedangkan parameter dan konfigurasi model disajikan pada Tabel 1:

Tabel 1. Parameter dan Konfigurasi Model

No	Parameter	Nilai / Keterangan
1	Model NLP	<i>Bidirectional Encoder Representations from Transformers (BERT) Base</i>
2	<i>Tokenizer</i>	<i>BertTokenizer</i>
3	Panjang maksimum token	128
4	Output embedding	<i>Token [CLS]</i> (lapisan terakhir)
5	Arsitektur pelatihan	<i>BertForSequenceClassification</i>
6	Mode pelatihan	<i>Regresi (num_labels = 1)</i>
7	<i>Optimizer</i>	<i>AdamW</i>
8	<i>Learning rate</i>	2e-5
9	<i>Epoch</i>	2
10	Rasio data latih : data uji	70%:30%
11	Metode kemiripan	<i>Cosine Similarity</i>
12	Metode evaluasi utama	<i>Mean Squared Error (MSE)</i>



Gambar 1 Alur Proses Penelitian Sistem Rekomendasi Musik

Data penelitian berupa 10.000 lirik lagu berbahasa Indonesia yang diperoleh dari *platform Kaggle*. Data diproses menggunakan bahasa pemrograman *Python* pada *Google Colaboratory*. Tahap pembersihan data dilakukan untuk menghapus data duplikat, lirik tidak lengkap, dan kesalahan format. Selanjutnya dilakukan

prapemrosesan teks yang meliputi pembersihan simbol dan tanda baca, *stopword removal*, tokenisasi, serta *stemming* menggunakan pustaka *Sastrawi* untuk menghasilkan teks yang lebih konsisten [7].

Representasi teks lirik dibentuk menggunakan model *Bidirectional Encoder Representations from Transformers (BERT) Base*. Model dilatih menggunakan arsitektur *BertForSequenceClassification* dalam mode regresi (*num_labels = 1*) [8]. Penggunaan pendekatan regresi dilakukan untuk memetakan representasi semantik lirik ke dalam skor numerik kontinu, yang selanjutnya digunakan sebagai dasar perhitungan kemiripan antar lagu. Pendekatan ini relevan dalam sistem rekomendasi berbasis *content-based filtering* karena memungkinkan pengukuran tingkat kesamaan secara lebih fleksibel dibandingkan klasifikasi diskret [9].

Embedding lirik diperoleh dari token *[CLS]* pada lapisan terakhir *BERT* karena token tersebut merepresentasikan makna keseluruhan teks. Dataset dibagi menjadi 70% data latih dan 30% data uji. Proses pelatihan menggunakan *optimizer AdamW* dengan learning rate sebesar $2e-5$ selama dua *epoch* [10].

Sistem rekomendasi dibangun menggunakan pendekatan *content-based filtering* dengan menghitung nilai *cosine similarity* antar embedding lirik untuk menentukan tingkat kemiripan antar lagu. Evaluasi kinerja model dilakukan menggunakan metrik *Mean Squared Error (MSE)* untuk menilai kesalahan prediksi skor regresi. Secara matematis, *MSE* dirumuskan sebagai berikut:

$$MSE = \left(\frac{1}{n}\right) \sum (y_i - \hat{y}_i)^2 \quad (1)$$

di mana y_i merupakan nilai aktual, $\hat{y}_i$ merupakan nilai prediksi, dan n adalah jumlah data uji.

Penggunaan *MSE* dalam penelitian ini bertujuan untuk mengukur sejauh mana kesalahan model dalam memetakan representasi semantik lirik ke skor numerik yang digunakan dalam proses rekomendasi berbasis kemiripan. Selain itu, evaluasi klasifikasi tambahan dilakukan dengan mengonversi skor regresi ke dalam kelas preferensi untuk mengukur nilai akurasi sebagai indikator pendukung kinerja sistem secara keseluruhan.

3. HASIL DAN ANALISIS

Bagian ini menyajikan hasil pengembangan dan evaluasi sistem rekomendasi musik berbasis analisis lirik menggunakan model *Bidirectional Encoder Representations from Transformers* dan pendekatan *Content-Based Filtering*. Analisis dilakukan mulai dari tahap pengolahan data hingga evaluasi performa sistem untuk memastikan validitas dan kontribusi empiris penelitian.

Dataset penelitian diperoleh dari *platform Kaggle* dan terdiri atas 10.000 lirik lagu berbahasa Indonesia. Setelah proses pembersihan data, sebanyak 1.458 data dihapus karena duplikasi, nilai kosong, dan campuran bahasa, sehingga dataset akhir berjumlah 8.542 lagu. Tahapan *data cleaning* ini penting karena kualitas data teks sangat memengaruhi stabilitas embedding dan kinerja model *Natural Language Processing (NLP)* [8].

Pra-pemrosesan teks dilakukan melalui normalisasi huruf, penghapusan simbol dan *stopword* Bahasa Indonesia, *tokenisasi*, serta *stemming*. Proses ini bertujuan mengurangi noise tanpa menghilangkan makna semantik utama lirik. Hasil pra-pemrosesan menunjukkan teks menjadi lebih ringkas dan konsisten, yang mendukung pembentukan representasi semantik yang lebih stabil pada model transformer [11]. Visualisasi hasil pembersihan data ditunjukkan pada Gambar 2.

	JUDUL LAGU	PENYANYI	ALBUM	GENRE	LIRIK LAGU (POTONGAN)	TAHUN RILIS
0	Bintang Luka	Yudha Halim	Hali Journey (2023)	Jazz	Ku tetap langit malam dan ingat senyummu. Masi...	2023
1	Selamanya Langit Menanti	Fajar Fadilah	Akhir Hits (2005)	Hip-Hop	Angin membawa kabar tentang rinduku. Jejak lan...	2005
2	Kota Hati Kisah	Bima Anggraini	Senja Journey (2021)	Hip-Hop	Kita menulis ulang cerita yang hampir pudar. D...	2021
3	Sendiri Kola	Ayu Kirana	Hali Story (2010)	EDM	Kau bawa pelangi setelah hari yang kelabu	2010
4	Kenangan Angin Langit	Nabila Saputra	Suara Notes (2002)	EDM	Angin membawa kabar tentang rinduku	2002
5	Langit Kisah	Bima Kirana	Sendin Notes (2020)	EDM	Rintik hujan menan di jendela rumah tua. Suar...	2020
6	Waktu Hujan	Intan Halim	Waktu Journey (2014)	Hip Hop	Ada tawa yang tersisa di kotak kenangan. Dinta...	2014
7	Waktu	Lina Kirana	Malam Journey (2008)	Pop	Hatiku sederhana namun selalu rindu. Angin mem...	2008
8	Akhir	Nabila Santoso	Mimpi Memories (2013)	Jazz	Kau bawa pelangi setelah han yang kelabu	2013
9	Jalan Selamanya	Fajar Kusuma	Bahagia Hits (2008)	Pop	Hatiku sederhana namun selalu rindu. Hatiku se...	2008

Gambar 2 Pembersihan Data

Representasi lirik lagu dibentuk menggunakan model *Bidirectional Encoder Representations from Transformers* dengan arsitektur *BertForSequenceClassification* dalam skema regresi (*num_labels = 1*). Penggunaan pendekatan regresi bertujuan memetakan hubungan semantik antar lirik ke dalam nilai numerik kontinu yang selanjutnya dimanfaatkan dalam penghitungan kemiripan. Strategi ini sejalan dengan penelitian

sebelumnya yang memanfaatkan regresi *BERT* untuk representasi semantik teks [12]. Hasil pelatihan menunjukkan penurunan nilai *loss* dari 1,4195 pada *epoch* pertama menjadi 1,3628 pada *epoch* kedua. Penurunan ini mengindikasikan bahwa model mampu mempelajari pola data, meskipun konvergensi yang terbatas menunjukkan perlunya *fine-tuning* lanjutan atau penambahan jumlah data pelatihan. Temuan ini konsisten dengan penelitian [9] yang menyatakan bahwa model *BERT* pada tugas regresi memerlukan optimasi lebih lanjut agar performanya lebih stabil yang ditampilkan pada Gambar 3.

```
train_df, test_df = train_test_split(df, test_size=0.30, random_state=42)
print("Train:", len(train_df))
print("Test :", len(test_df))

• Train: 5979
  Test : 2563
```

Gambar 3 Data Hasil Train dan Test Embedding Berbasis *BERT*

Embedding lirik yang dihasilkan kemudian digunakan dalam sistem rekomendasi berbasis konten dengan metode *cosine similarity*. Hasil pengujian menunjukkan nilai kemiripan antarlagu berada pada rentang 0,989 hingga 0,994 terhadap lagu acuan. Nilai yang sangat tinggi ini menunjukkan bahwa embedding berbasis *BERT* mampu menangkap kesamaan tema, emosi, dan konteks linguistik lirik secara efektif [13]. Namun demikian, tingkat kemiripan yang mendekati 1 juga mengindikasikan potensi *over-similarity*, yaitu kondisi ketika banyak lagu dianggap sangat mirip sehingga dapat mengurangi keberagaman rekomendasi yang menyarankan integrasi fitur tambahan seperti metadata atau fitur audio untuk meningkatkan diversitas hasil rekomendasi. pada Gambar 4.

```
▼ ... Training Epoch 1: 100%| 748/748 [2:21:50<00:00, 11.38s/it]
Loss Epoch 1: 1.4195148697431712
Training Epoch 2: 100%| 748/748 [2:21:04<00:00, 11.32s/it] Loss Epoch 2: 1.3628468776450438
```

Gambar 4 Nilai Loss Model *BERT* (Regresi)

Evaluasi performa dilakukan menggunakan *Mean Squared Error (MSE)* sebagai metrik utama untuk menilai akurasi prediksi regresi. Hasil pengujian menunjukkan nilai *MSE* sebesar 1,4066, yang mengindikasikan bahwa rata-rata kesalahan prediksi masih relatif tinggi. Kondisi ini menunjukkan bahwa pemetaan hubungan semantik lirik ke skor numerik belum sepenuhnya optimal, sebagaimana juga ditemukan pada penelitian rekomendasi berbasis teks murni [14]. Selain itu, evaluasi klasifikasi tambahan dilakukan dengan mengonversi skor regresi ke dalam tiga kelas preferensi. Hasil menunjukkan nilai akurasi sebesar 0,8369, namun nilai *precision* yang rendah (0,1219) mengindikasikan adanya ketidakseimbangan kelas prediksi. Hal ini menegaskan bahwa nilai akurasi yang tinggi tidak selalu mencerminkan performa klasifikasi yang baik, terutama pada dataset dengan distribusi kelas yang tidak seimbang. Hasil pengujian menunjukkan nilai *MSE* sebesar 1,4066 sebagaimana ditampilkan pada Gambar 5.

```
=====
MSE : 1.4066150190493631
=====
```

Gambar 5 Data Hasil *Mean Squared Error (MSE)*

Dibandingkan dengan pendekatan berbasis *TF-IDF* atau *Word2Vec*, penggunaan embedding kontekstual *BERT* dalam penelitian ini mampu menangkap makna lirik secara lebih mendalam dan menghasilkan nilai kemiripan yang lebih konsisten. Penelitian sebelumnya menunjukkan bahwa metode klasik cenderung gagal merepresentasikan konteks semantik yang kompleks dalam lirik lagu [15]. Kontribusi utama penelitian ini terletak pada penerapan embedding kontekstual *BERT* dalam sistem rekomendasi musik berbasis lirik berbahasa Indonesia serta analisis empiris kinerjanya menggunakan metrik regresi dan klasifikasi. Meskipun demikian, hasil evaluasi menunjukkan bahwa pengembangan lanjutan masih diperlukan, khususnya untuk meningkatkan stabilitas regresi dan keberagaman rekomendasi [13].

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem rekomendasi musik berbasis analisis semantik lirik lagu menggunakan pendekatan *Natural Language Processing* dengan model *Bidirectional Encoder Representations from Transformers* dan metode *Content-Based Filtering*. Hasil eksperimen menunjukkan bahwa embedding lirik berbasis *BERT* mampu merepresentasikan makna semantik secara kontekstual, yang tercermin dari nilai *cosine similarity* yang tinggi (0,989–0,994) pada lagu-lagu yang direkomendasikan. Evaluasi kinerja sistem

menghasilkan nilai *Mean Squared Error* sebesar 1,4066 serta akurasi klasifikasi sebesar 0,8369, meskipun nilai *precision* masih rendah akibat ketidakseimbangan kelas. Temuan ini menegaskan kontribusi ilmiah utama penelitian berupa pemanfaatan embedding kontekstual *BERT* untuk lirik lagu berbahasa Indonesia dalam sistem rekomendasi musik berbasis konten, yang terbukti lebih efektif dibandingkan pendekatan berbasis representasi teks konvensional. Secara praktis, model yang diusulkan berpotensi meningkatkan relevansi dan personalisasi rekomendasi musik pada platform digital, meskipun pengembangan lanjutan seperti *fine-tuning* model, penyeimbangan data, serta integrasi fitur metadata dan audio masih diperlukan untuk meningkatkan stabilitas dan akurasi sistem secara keseluruhan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika, Fakultas Teknik, Universitas Ma'arif Nahdlatul Ulama Kebumen atas dukungan fasilitas dan lingkungan akademik yang mendukung terlaksananya penelitian ini. Ucapan terima kasih juga disampaikan kepada seluruh pihak yang telah memberikan masukan dan bantuan selama proses penelitian dan penyusunan artikel ilmiah ini.

REFERENSI

- [1] Z. Li *et al.*, "Graph and Sequential Neural Networks in Session-based Recommendation: A Survey," vol. 57, no. 2, 2024, doi: 10.1145/3696413.
- [2] E. Oramas, S., Brost, B., & Gómez, "Lyric-based music embedding for music recommendation.," 2021, doi: <https://doi.org/10.5334/tismir.55>.
- [3] R. Kim, H., & Akama, "A computational analysis of lyric similarity perception.," 2024, doi: <https://arxiv.org/abs/2404.02342>.
- [4] X. Wang, H., Huang, J., & Yang, "Transformer-based neural recommender systems: A survey," 2023, doi: <https://doi.org/10.1145/3593055>.
- [5] R. B. Daniel Kostrzewa, Jonatan Chrobak, "Attributes Relevance in Content-Based Music Recommendation System," 2024, doi: <https://doi.org/10.3390/app14020855>.
- [6] R. Liyanarachchi, A. Joshi, and E. Meijering, "A Survey on Multimodal Music Emotion Recognition," *ACM Comput. Surv.*, vol. ?, no. ?, 2025, [Online]. Available: <http://arxiv.org/abs/2504.18799>
- [7] Emilia Parada-Cabaleiro, "Song lyrics have become simpler and more repetitive over the last five decades," 2024, doi: 10.1038/s41598-024-55742-x.
- [8] M. Sahel Muhyidin, I. Hariyanti, M. F. Novianto, A. A. Apriliani, and S. N. Rohmah, "Analisis Sistem Rekomendasi Musik Berdasarkan Lirik Dengan Metode Term Frequency-Inverse Document Frequency (Tf-Idf)," *Semin. Teknol. Majalengka*, vol. 8, 2024.
- [9] V. R. Joseph, "No Title," *Optim. ratio data Split. Stat. Anal. Data Min. ASA Data Sci. Journal*, 15(6), 531–534., 2022, doi: <https://doi.org/10.1002/sam.11583>.
- [10] G. Manueke, "Analisis Sentimen Lirik Lagu Populer dengan BERT," 2023, doi: 10.33830/jurnal.v4i1.61.
- [11] S. N. R. Muhammad Sahel Muhyidin, I Fani Hariyanti, Muhammad Fahmi Novianto, Anna Alifia Apriliani, "Analisis Sistem Rekomendasi Musik Berdasarkan Lirik Dengan Metode Term Frequency-Inverse Document Frequency," 2024, [Online]. Available: https://prosiding.unma.ac.id/index.php/stima/article/view/1303?utm_source=chatgpt.com
- [12] Yongdong Zhang, *Music Information Retrieval and Recommendation*. Wiley, 2023. doi: 10.1002/9781119822109.
- [13] S. Kumar and P. Jadon, "Machine Learning Techniques and Advancement in Recommendation System," *SSRN Electron. J.*, 2024, doi: 10.2139/ssrn.4934448.
- [14] Reyhan Alana, "Implementasi Algoritma Content Based Filtering dalam Sistem Rekomendasi Komik," 2024, doi: 10.32520/stmsi.v13i4.1344.
- [15] H. D. (2025). Ningrum, M. P., Mutia, R., Azmi, H., & Khalifah, "Sentiment analysis of Twitter reviews on Google Play Store using a combination of convolutional neural network and long short-term memory algorithms. *Public Research Journal of Engineering, Data Technology and Computer Science*, 2(2), 107–115.," 2025, doi: <https://doi.org/10.57152/predatecs.v2i2.1625>.