Prosiding Seminar Nasional Teknologi dan Sistem Informasi (SITASI) 2026
Surabaya, 3 Juni 2026

KOMPARASI LDA DAN BERTOPIC UNTUK PEMODELAN TOPIK PADA ULASAN APLIKASI IKD

COMPARISON OF LDA AND BERTOPIC FOR TOPIC MODELING ON IKD APPLICATION REVIEWS

Viviana Purba*, Eka Dyar Wahyuni[2], Tri Luhur Indayanti Sugata[3]
*E-mail: 22082010152@student.upnjatim.ac.id

[1,2,3]Sistem Informasi, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur

Abstrak

Aplikasi Identitas Kependudukan Digital (IKD) merupakan bagian dari transformasi layanan administrasi publik di Indonesia yang memungkinkan masyarakat mengakses dokumen kependudukan secara digital. Seiring meningkatnya jumlah pengguna, ulasan pada Google Play Store menjadi sumber informasi penting untuk memahami pengalaman dan permasalahan pengguna. Namun, banyaknya volume ulasan dengan bentuk data teks yang tidak terstruktur menyebabkan analisis manual menjadi kurang efektif. Penelitian ini bertujuan membandingkan metode Latent Dirichlet Allocation (LDA) dan BERTopic dalam pemodelan topik pada ulasan aplikasi IKD. Data penelitian diperoleh melalui proses scraping Google Play Store dengan total 13.197 ulasan, kemudian melalui tahap pembersihan, filtering, dan preprocessing hingga diperoleh 6.891 ulasan relevan. Evaluasi model dilakukan menggunakan metrik coherence score dan topic diversity. Hasil penelitian menunjukkan bahwa model LDA terbaik menghasilkan 3 topik dengan coherence score sebesar 0,5083 dan topic diversity sebesar 0,9333. Sementara itu, BERTopic menghasilkan performa yang lebih baik dengan 5 topik pasca-topic merging, coherence score sebesar 0,6196, dan topic diversity sebesar 0,92. Topik yang dihasilkan BERTopic juga lebih spesifik dan representatif, seperti kendala akses akun, proses aktivasi barcode Dukcapil, verifikasi foto wajah, layanan dokumen digital, dan kompatibilitas perangkat Android. Hasil penelitian menunjukkan bahwa BERTopic lebih efektif dibandingkan LDA dalam menangkap konteks semantik pada ulasan pendek dan tidak terstruktur, sehingga lebih optimal digunakan untuk analisis topik pada ulasan aplikasi layanan publik digital.

Kata kunci**:** *topic modeling, LDA, BERTopic, IKD, coherence score, diversity*

Abstract [Times New Roman 11 Normal Justify]

The Identitas Kependudukan Digital (IKD) application is part of the digital transformation of public administrative services in Indonesia, enabling citizens to access population administration documents digitally. As the number of users continues to increase, reviews on Google Play Store have become an important source of information for understanding user experiences and issues. However, the large volume of reviews in the form of unstructured text data makes manual analysis less effective. This study aims to compare the Latent Dirichlet Allocation (LDA) and BERTopic methods for topic modeling on IKD application reviews. The research data were collected through Google Play Store scraping, resulting in a total of 13,197 reviews, which then underwent cleaning, filtering, and preprocessing stages to obtain 6,891 relevant reviews. Model evaluation was conducted using coherence score and topic diversity metrics. The results show that the best LDA model generated 3 topics with a coherence score of 0.5083 and a topic diversity score of 0.9333. Meanwhile, BERTopic achieved better performance with 5 topics after topic merging, a coherence score of 0.6196, and a topic diversity score of 0.92. The topics generated by BERTopic were also more specific and representative, including account access issues, Dukcapil barcode activation

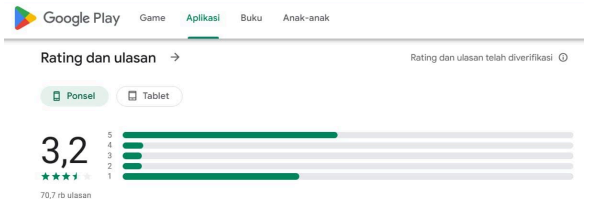
processes, facial verification, digital document services, and Android device compatibility. The findings indicate that BERTopic is more effective than LDA in capturing semantic context within short and unstructured reviews, making it more suitable for topic analysis of digital public service application reviews.

***Keywords**: topic modeling, LDA, BERTopic, IKD, coherence score, diversity.*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi yang semakin pesat telah membawa perubahan pada berbagai aspek kehidupan masyarakat, termasuk dalam bidang pelayanan administrasi publik. Transformasi digital mendorong instansi pemerintah untuk mengubah sistem pelayanan konvensional menjadi layanan berbasis elektronik yang lebih efisien, transparan, dan responsif terhadap kebutuhan masyarakat[1]. Perubahan ini tidak lagi hanya menjadi pilihan, tetapi juga menjadi kebutuhan agar birokrasi publik mampu mengikuti perkembangan masyarakat yang semakin dinamis dan kompleks. Oleh karena itu, berbagai instansi pemerintah, baik di tingkat pusat maupun daerah, terus berupaya mengembangkan inovasi layanan digital untuk meningkatkan kualitas pelayanan publik.

Salah satu bentuk nyata dari digitalisasi layanan administrasi publik di Indonesia adalah peluncuran aplikasi Identitas Kependudukan Digital (IKD) oleh Direktorat Jenderal Kependudukan dan Pencatatan Sipil Kementerian Dalam Negeri pada tahun 2022[2]. Aplikasi ini memungkinkan masyarakat mengakses dokumen kependudukan secara digital melalui perangkat gawai sehingga proses verifikasi identitas menjadi lebih praktis tanpa perlu membawa dokumen fisik. Implementasi IKD yang dilakukan secara bertahap di berbagai daerah menjadi langkah strategis dalam mendukung transformasi penggunaan KTP elektronik (e-KTP) menuju identitas digital[3]. Hingga saat ini, aplikasi IKD telah diunduh lebih dari 10 juta kali di Google Play Store, yang menunjukkan tingginya tingkat adopsi masyarakat serta potensi pengembangan layanan digital kependudukan di masa mendatang.



Gambar 1. Rating Aplikasi IKD di Google Playstore

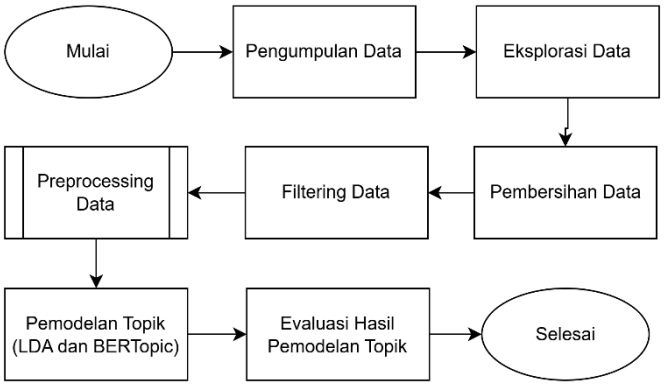
Sebagaimana ditunjukkan pada Gambar 1, IKD telah mendapatkan nilai rating 3,2 dari sekitar 70,7 ribu ulasan dengan penilaian yang bervariasi mulai dari bintang satu hingga bintang lima (Desember, 2025). Ulasan tersebut dapat dimanfaatkan untuk mengidentifikasi permasalahan, mengevaluasi performa aplikasi, serta menentukan strategi pengembangan fitur yang lebih sesuai dengan kebutuhan pengguna[4]. Namun, jumlah ulasan yang besar dan beragam, serta berbentuk teks tidak terstruktur, sering kali menyulitkan proses analisis secara manual. Dalam konteks ini, pemodelan topik *(topic modeling*) menjadi pendekatan yang relevan untuk digunakan. Dengan pemodelan topik, tema-tema tersembunyi dari ribuan hingga jutaan ulasan yang tidak terstruktur dapat diidentifikasi secara otomatis[5]. Pemodelan topik memungkinkan memahami aspek apa saja yang paling banyak dibahas oleh pengguna seperti fitur, kendala teknis, kemudahan penggunaan, maupun layanan. Dengan demikian, teknik ini membantu mengorganisasi data teks ulasan yang beragam menjadi representasi topik yang lebih terstruktur dan informatif.

Berbagai metode telah dikembangkan untuk melakukan pemodelan topik, di antaranya adalah Latent Dirichlet Allocation (LDA) dan BERTopic[6]. LDA merupakan metode probabilistik yang telah lama digunakan dan sering dijadikan sebagai *baseline* dalam penelitian pemodelan topik karena kesederhanaan serta kemampuannya dalam mengidentifikasi distribusi topik dalam

dokumen. Terlepas dari popularitasnya, LDA memiliki keterbatasan dalam mengolah dokumen pendek dan tidak terstruktur, yang sering kali memicu terbentuknya topik redundan atau kurang spesifik. Kendala lainnya terletak pada sifat model yang tidak deterministik, sehingga menghasilkan output yang bervariasi meskipun dijalankan pada dataset yang sama[6]. Di sisi lain, BERTopic merupakan pendekatan yang lebih modern dengan memanfaatkan representasi berbasis transformer seperti BERT, sehingga mampu menangkap makna semantik yang lebih mendalam dalam teks[8]. Keunggulan ini menjadikan BERTopic lebih adaptif dalam menangani teks ulasan yang umumnya bersifat pendek, tidak terstruktur, dan kaya konteks. Oleh karena itu, penelitian ini melakukan komparasi antara metode LDA dan BERTopic untuk pemodelan topik pada ulasan aplikasi IKD. Perbandingan ini bertujuan untuk mengevaluasi kemampuan masing-masing metode dalam mengidentifikasi topik yang representatif, serta menentukan pendekatan yang lebih efektif dalam mengekstraksi informasi dari ulasan pengguna aplikasi Identitas Kependudukan Digital.

2. METODOLOGI

Gaambar 2 adalah metode penelitian yang digunakan dalam penelitian ini.



Gambar 2. Alur Penelitian[6]

2.1 Pengumpulan Data

Data penelitian diperoleh dari ulasan pengguna aplikasi Identitas Kependudukan Digital (IKD) yang tersedia pada platform Google Play Store. Proses pengumpulan data dilakukan melalui teknik web scraping menggunakan google-play-scraper, yang memungkinkan pengambilan data secara otomatis dan terstruktur. Waktu pengambilan data menggunakan variabel startDate (1 Juni 2024) dan endDate (30 November 2025). Rentang waktu ini dipilih dengan mempertimbangkan periode mulai diberlakukannya kewajiban penggunaan aplikasi IKD di Indonesia serta kesesuaiannya dengan periode penelitian

2.2 Eksplorasi Data

Setelah proses pengumpulan selesai, data terlebih dahulu dieksplorasi untuk memahami karakteristik dasarnya sebelum memasuki tahapan analisis lanjutan dan pemodelan topik. Eksplorasi ini mencakup pemeriksaan jumlah baris dan kolom, identifikasi tipe data pada setiap atribut, pengecekan keberadaan nilai kosong, dan pengecekan distribusi data. Langkah awal ini penting untuk memberikan gambaran umum mengenai kondisi dataset dan memastikan bahwa data berada dalam keadaan yang layak untuk diproses lebih lanjut.

2.3 Pembersihan Data

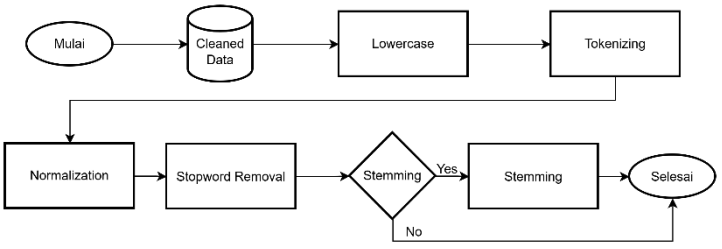
Pembersihan data dilakukan untuk mengurangi noise pada data serta mengatasi permasalahan seperti format data yang tidak konsisten, nilai kosong, dan duplikasi, sehingga dapat mencegah bias dan meningkatkan keandalan hasil analisis. Pada tahap ini, dilakukan penghapusan kolom-kolom yang tidak digunakan serta penggantian nama kolom content menjadi kolom ulasan agar sesuai dengan konteks analisis. Selain itu, dilakukan penghapusan data duplikat untuk memastikan setiap

ulasan bersifat unik. Proses pembersihan teks juga dilakukan dengan menghapus mentions, angka, URL, serta karakter non-alfanumerik yang tidak relevan.

2.4 Filtering Data

Filtering data dilakukan untuk memastikan bahwa ulasan yang dianalisis benar-benar relevan dan memiliki informasi yang memadai. Pada tahap ini, data hasil pembersihan sebelumnya diseleksi dengan menghapus baris ulasan yang memiliki jumlah kata kurang dari dua kata karena dianggap tidak mengandung informasi yang cukup untuk dianalisis lebih lanjut. Selain itu, dilakukan pengecekan dan penghapusan ulasan yang tidak relevan, yaitu ulasan yang tidak berkaitan dengan aplikasi Identitas Kependudukan Digital (IKD).

2.5 Preprocessing Data



Gambar 3. Tahapan Preprocessing

Sebagaimana ditunjukkan pada Gambar 3, tahapan yang dilakukan pada proses preprocessing meliputi lowercase, tokenizing, normalization, stopwords removal, dan stemming.

2.5.1 Lowercase

Lowercase adalah proses mengubah seluruh karakter dalam teks menjadi huruf kecil. Tujuannya adalah untuk menghindari perbedaan makna antara kata yang sama tetapi memiliki kapitalisasi berbeda, sehingga model tidak memperlakukan “Aplikasi” dan “aplikasi” sebagai dua token yang berbeda.

2.5.1 *Tokenizing*

Tokenizing adalah proses memecah teks menjadi unit-unit kecil berupa token yang biasanya kata atau simbol tertentu. Proses ini mempermudah analisis lebih lanjut karena model akan memproses teks per kata, bukan per kalimat secara utuh.

2.5.1 *Normalization*

Normalization adalah proses mengubah kata-kata tidak baku, slang, atau variasi penulisan menjadi bentuk baku. Tujuan dari normalisasi adalah untuk mengurangi variasi kata yang memiliki makna sama, sehingga meningkatkan konsistensi data dan membantu model dalam mengenali pola kata,terutama pada teks ulasan yang banyak menggunakan bahasa informal.

2.5.1 *Stopword Removal*

Stopword removal adalah proses menghapus kata-kata umum yang tidak memiliki makna signifikan dalam konteks analisis, seperti “dan”, “yang”, atau “atau”. Penghapusan stopword dilakukan untuk mengurangi noise dan memastikan bahwa model hanya memproses kata-kata yang memiliki nilai informasi penting.

2.5.1 *Stemming*

Stemming adalah proses mengubah kata-kata berimbuhan menjadi bentuk dasarnya dengan cara menghilangkan imbuhan seperti awalan, akhiran, sisipan, atau gabungan imbuhan tersebut. Proses stemming membantu mengelompokkan berbagai bentuk kata yang secara makna sama ke akar kata yang seragam, sehingga mengurangi variasi kata dalam data teks.

2.6 Pemodelan Topik

2.6.1 LDA

Latent Dirichlet Allocation (LDA) adalah model hierarki Bayesian tiga tingkat yang mengasumsikan setiap dokumen sebagai perpaduan dari sekumpulan topik tertentu, di mana setiap topik tersebut didefinisikan sebagai distribusi probabilitas atas kumpulan kata-kata[7]. Pemodelan

topik LDA dilakukan menggunakan data hasil preprocessing dengan stemming dan library gensim. Tahapan meliputi pembentukan dictionary dan corpus, kemudian pelatihan model untuk menghasilkan distribusi topik. Parameter seperti jumlah topik, passes, dan alpha dioptimasi menggunakan grid search berdasarkan nilai *coherence score* tertinggi.

2.5.1 BERTopic

Pemodelan BERTopic menggunakan data tanpa stemming. Proses diawali dengan pembuatan document embedding menggunakan SentenceTransformer berbasis BERT, dilanjutkan reduksi dimensi dengan UMAP, dan clustering menggunakan HDBSCAN. Representasi topik kemudian dihasilkan menggunakan class-based TF-IDF (c-TF-IDF) dengan menggabungkan seluruh dokumen dalam satu klaster dan menghitung kata-kata yang paling menonjol untuk mendefinisikan setiap topik[5]

2.7 Evaluasi Hasil Pemodelan Topik

Evaluasi kualitas topic modeling tidak cukup hanya menggunakan coherence, tetapi juga perlu mempertimbangkan diversity untuk memastikan topik yang dihasilkan tidak redundan dan tetap memiliki keberagaman yang baik[9]. Secara umum, nilai coherence score yang semakin tinggi menunjukkan bahwa topik yang dihasilkan memiliki keterkaitan semantik yang lebih konsisten dan lebih mudah dipahami oleh manusia. Topik dengan nilai tinggi umumnya tersusun dari kata-kata yang sering muncul bersama dalam konteks yang serupa, sehingga membentuk suatu konsep yang lebih jelas dan bermakna. Sebaliknya, nilai coherence score yang rendah mengindikasikan bahwa kata-kata dalam topik memiliki hubungan semantik yang lemah, sehingga topik menjadi kurang jelas, sulit diinterpretasikan, atau tidak merepresentasikan tema tertentu secara baik[10]. Lebih lanjut, diversity berperan penting dalam mencegah terjadinya redundansi antar topik, di mana beberapa topik memiliki kata kunci yang mirip atau identik. Nilai diversity yang tinggi menunjukkan bahwa model mampu menghasilkan topik yang beragam dan mencakup berbagai aspek dalam data[11].

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Sebanyak 13.197 ulasan berhasil dikumpulkan dari pengguna aplikasi Identitas Kependudukan Digital (IKD) di Google Play Store melalui metode web scraping menggunakan library google-play-scraper. Proses ini memungkinkan pengambilan data secara otomatis dan terstruktur. Data yang diperoleh terdiri dari 11 atribut, yaitu reviewId, userName, userImage, content, score, thumbsUpCount, reviewCreatedVersion, at, replyContent, repliedAt, dan appVersion.

```
df_ikd = pd.DataFrame(ikd_googleplaystore_reviews)
print("Shape:", df_ikd.shape)
df_ikd.sample(7)
```

Shape: (13197, 11)

	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion
843	d6f0b31f-9c5a-4005-b41a-ec6f2c45f7ce	ilah kamilah	https://play-lh.googleusercontent.com/a/ACg8oc...	sudah jelas mau mendaftar online tapi diminta ...	1	0	2.0.1	2025-10-26 12:04:59	None	None	2.0.1
1276	8365367e-19d3-4f9c-b4f8-594e2abe3d0d	Suara Jaka	https://play-lh.googleusercontent.com/a-/ALV-U...	disdukcapil kota balam pelayanan bagus	5	0	2.0.1	2025-10-02 03:06:09	None	None	2.0.1
8165	e4062196-6fc7-418b-8e7b-916868ecf644	poer java	https://play-lh.googleusercontent.com/a-/ALV-U...	Aplikasi kayak asal asalan bikinya...ga guna	1	0	2.0.1	2025-01-14 16:17:08	None	None	2.0.1
8065	71aaec16-3cf2-4e98-916b-cbab523249b7	Filemon Koesnoto	https://play-lh.googleusercontent.com/a-/ALV-U...	Cerdas dn canggih srhua sdh dilangan tanpa hrs ...	5	0	2.0.1	2025-01-15 13:22:29	None	None	2.0.1
13001	c174f3a5-6536-43c1-85d2-343339efb1d	Jody Motor23	https://play-lh.googleusercontent.com/a/ACg8oc...	Termakasih Capil balam udah membuat kan ikd jl...	5	0	1.2.2	2024-06-10 04:58:52	None	None	1.2.2
7766	d02459c9-a7c3-4ecb-84ac-4e3a4d65c767	Dian Hasbullah	https://play-lh.googleusercontent.com/a/ACg8oc...	Selalu dan selalu aplikasi pemerintahan buruk	1	0	2.0.1	2025-01-19 16:22:44	None	None	2.0.1
5467	a4ede2e3-ce56-4494-8488-eafd2b984833	Wardin	https://play-lh.googleusercontent.com/a-/ALV-U...	good	5	0	2.0.1	2025-03-25 18:18:25	None	None	2.0.1

Gambar 4. Hasil Scrapping Data Ulasan IKD

3.2 Eksplorasi Data

Setelah diekplorasi, data menunjukkan bahwa mayoritas ulasan berada pada rating 1 (sangat rendah) dengan jumlah yang jauh lebih dominan dibandingkan rating lainnya, diikuti oleh rating 5 sebagai jumlah terbesar kedua. Pola ini mengindikasikan bahwa pengalaman pengguna terhadap aplikasi IKD cenderung bersifat polarized, yaitu didominasi oleh ulasan sangat negatif dan sebagian ulasan sangat positif. Temuan ini menjadi indikasi awal bahwa analisis emosi diperlukan untuk menggali lebih dalam alasan di balik persepsi negatif maupun positif pengguna.



Gambar 5. *Wordcloud* Ulasan

Selanjutnya dilakukan eksplorasi pada kolom content untuk mengetahui karakteristik awal teks ulasan pengguna. Visualisasi word cloud pada Gambar 5 menunjukkan bahwa dataset ulasan masih didominasi oleh kata-kata umum seperti 'dan', 'nya', dan 'saya' yang tidak memiliki nilai informatif dalam analisis emosi maupun aspek. Meskipun demikian, terdapat pola kemunculan kata kunci seperti 'tidak bisa' dan 'ribet' yang memberikan indikasi awal mengenai keluhan pengguna. Hasil visualisasi ini memperkuat perlunya dilakukan proses preprocessing yang lebih intensif guna mengekstraksi kata-kata bermakna yang dapat merepresentasikan topik utama dan emosi pengguna secara tepat.

3.3 Pembersihan Data

Proses ini diawali dengan menghapus atribut yang tidak relevan terhadap tujuan penelitian. Kolom-kolom seperti reviewId, userName, userImage, thumbsUpCount, reviewCreatedVersion, at, replyContent, repliedAt, appVersion, dan score dihapus karena tidak digunakan dalam penelitian analisis teks ini. Hasil pengecekan menunjukkan adanya data duplikat, sehingga dilakukan penghapusan menggunakan fungsi drop_duplicates() untuk menghindari bias dalam analisis, terutama pada tahap pemodelan topik. Setelah itu dilakukan pembersihan teks ulasan menggunakan fungsi khusus clean_review_text yang dirancang untuk menghapus elemen-elemen yang tidak relevan dalam analisis teks, seperti mention URL, angka, serta karakter non-alfanumerik.

3.4 Filtering Data

Proses filtering dilakukan dengan menghapus ulasan yang terdiri dari kurang dari dua kata, karena dianggap tidak mengandung informasi yang cukup untuk dianalisis. Selanjutnya, dilakukan validasi manual untuk memastikan relevansi ulasan terhadap konteks aplikasi IKD. Ulasan yang tidak relevan mencakup komentar di luar konteks aplikasi, ulasan yang terlalu umum tanpa informasi spesifik, serta ulasan dalam bahasa daerah. Tahapan ini menghasilkan dataset berisi 6891 ulasan yang telah melalui dua tahap penyaringan yang lebih bersih dan relevan melalui penyaringan otomatis dan manual.

3.5 Preprocessing Data

Tahap preprocessing dilakukan untuk mengubah data teks menjadi bentuk yang lebih terstruktur dan siap digunakan dalam proses pemodelan. Pada penelitian ini, preprocessing dilakukan melalui beberapa subproses yaitu lowercase, tokenizing, normalization, stopwords removal, dan stemming. Setiap tahapan dilakukan secara bertahap dan hasilnya ditampilkan untuk menunjukkan perubahan pada data. Hasil preprocessing dapat dilihat pada Gambar 6.

	ulasan	lowercase	tokens	normalized	no_stopwords	stemmed
3461	Abis apdet ko malah isi biodata lagi	abis apdet ko malah isi biodata lagi	[abis, apdet, ko, malah, isi, biodata, lagi]	[abis, apdet, ko, malah, isi, biodata, lagi]	[abis, apdet, ko, isi, biodata]	[abis, apdet, ko, isi, biodata]
8582	Sudah lumayan ada baiknya ada permohonan perubahan data seperti mau mengubah data golongan darah karena akan bermanfaat sebagai database golongan darah	sudah lumayan ada baiknya ada permohonan perubahan data seperti mau mengubah data golongan darah karena akan bermanfaat sebagai database golongan darah	[sudah, lumayan, ada, baiknya, ada, permohonan, perubahan, data, seperti, mau, mengubah, data, golongan, darah, karena, akan, bermanfaat, sebagai, database, golongan, darah]	[sudah, lumayan, ada, baiknya, ada, permohonan, perubahan, data, seperti, mau, menguben, data, golongan, darah, karena, akan, bermanfaat, sebagai, database, golongan, darah]	[lumayan, baiknya, permohonan, perubahan, data, menguben, data, golongan, darah, bermanfaat, database, golongan, darah]	[lumayan, baik, mohon, ubah, data, ubah, data, golong, darah, manfaat, database, golong, darah]
9041	aplikasi gak guna gak praktis mesa mau daftar aja harus pergi kekantor lah iya kalau tinggalnya di kota kalau dikampung masa harus kekota dulu	aplikasi gak guna gak praktis mesa mau daftar aja harus pergi kekantor lah iya kalau tinggalnya di kota kalau dikampung masa harus kekota dulu	[aplikasi, gak, guna, gak, praktis, mesa, mau, daftar, aja, harus, pergi, kekantor, lah, iya, kalau, tinggalnya, di, kota, kalau, dikampung, masa, harus, kekota, dulu]	[aplikasi, tidak, guna, tidak, praktis, mesa, mau, daftar, saja, harus, pergi, kekantor, lah, iya, kalau, tinggalnya, di, kota, kalau, dikampung, masa, harus, kekota, dulu]	[praktis, daftar, pergi, kekantor, iya, tinggalnya, kota, dikampung, kekota]	[praktis, daftar, pergi, kantor, iya, tinggal, kota, kampung, kota]
6906	Alhamdulillah cuma sekali verifikasi ke dukcapil manfaatnya buat seumur hidup pelayanan tinggal klik klik kelar	alhamdulillah cuma sekali verifikasi ke dukcapil manfaatnya buat seumur hidup pelayanan tinggal klik klik kelar	[alhamdulillah, cuma, sekali, verifikasi, ke, dukcapil, manfaatnya, buat, seumur, hidup, pelayanan, tinggal, klik, klik, kelar]	[alhamdulillah, cuma, sekali, verifikasi, ke, dukcapil, manfaatnya, buat, seumur, hidup, pelayanan, tinggal, klik, klik, kelar]	[alhamdulillah, verifikasi, dukcapil, manfaatnya, seumur, hidup, pelayanan, tinggal, klik, klik, kelar]	[alhamdulillah, verifikasi, dukcapil, manfaat, umur, hidup, layan, tinggal, klik, klik, kelar]
2232	kenapa eror terus bagian barcode	kenapa eror terus bagian barcode	[kenapa, eror, terus, bagian, barcode]	[kenapa, eror, terus, bagian, barcode]	[eror, barcode]	[eror, barcode]

Gambar 6. Hasil Preprocessing

3.6 Pemodelan Topik

3.6.1 LDA

Proses pemodelan topik dijalankan menggunakan algoritma Latent Dirichlet Allocation (LDA) melalui library Gensim. Tahap awal dimulai dengan pembentukan dictionary menggunakan fungsi corpora.Dictionary untuk memetakan kata unik dari seluruh dokumen. Selanjutnya, data ditransformasi menjadi format corpus melalui metode Bag-of-Words (BoW) dengan fungsi doc2bow(), yang merepresentasikan frekuensi kemunculan kata pada tiap dokumen. Untuk memperoleh jumlah topik yang paling optimal, dilakukan grid search pada rentang 2 hingga 10 topik menggunakan fungsi compute_coherence_values(). Evaluasi setiap model dilakukan dengan parameter passes=50 dan alpha='symmetric', serta diukur menggunakan metrik coherence score (c_v) untuk menilai keterkaitan semantik antar kata.

Table 1. Coherence Score LDA

Jumlah Topik	Coherence Score
2	0.4213
3	**0.5083**
4	0.4958
5	0.4637
6	0.4318
7	0.4658
8	0.4499
9	0.4575
10	0.4416

Model LDA paling optimal dipilih berdasarkan nilai coherence score tertinggi dari sekumpulan model dalam model_list. Berdasarkan hasil pengujian pada Tabel 1, ditetapkan bahwa jumlah topik paling ideal adalah tiga topik dengan skor koherensi sebesar 0,5083. Setelah model optimal terpilih, dilakukan evaluasi terhadap skor diversity serta ekstraksi kata kunci berbobot terbesar. Model ini menghasilkan skor diversity yang tinggi, yaitu 0,9333, yang menunjukkan perbedaan karakteristik antar-topik yang sangat jelas. Sebagai dasar tahap interpretasi, setiap topik direpresentasikan oleh 10 kata kunci utama yang rincian lengkapnya disajikan pada Tabel 2.

Table 2. Daftar Kata pada Topik LDA

Topik	Daftar Kata	Interpretasi
0	0.039*"daftar" + 0.035*"online" + 0.034*"dukcapil" + 0.023*"barcode" + 0.022*"kantor" + 0.021*"scan" + 0.017*"bikin" + 0.016*"qr" + 0.016*"tugas" + 0.013*"mudah"	Proses Registrasi dan Aktivasi
1	0.055*"masuk" + 0.031*"update" + 0.026*"foto" + 0.024*"daftar" + 0.021*"pin" + 0.018*"buka" + 0.017*"data" + 0.017*"ulang" + 0.014*"akun" + 0.013*"tolong"	Kendala Teknis dan Akses Akun
2	0.042*"ktp" + 0.020*"layan" + 0.019*"data" + 0.016*"kk" + 0.014*"digital" + 0.014*"cetak" + 0.013*"dokumen" + 0.012*"unduh" + 0.010*"d" + 0.009*"proses"	Fungsi Layanan Dokumen Digital

2.6.2 BERTopic

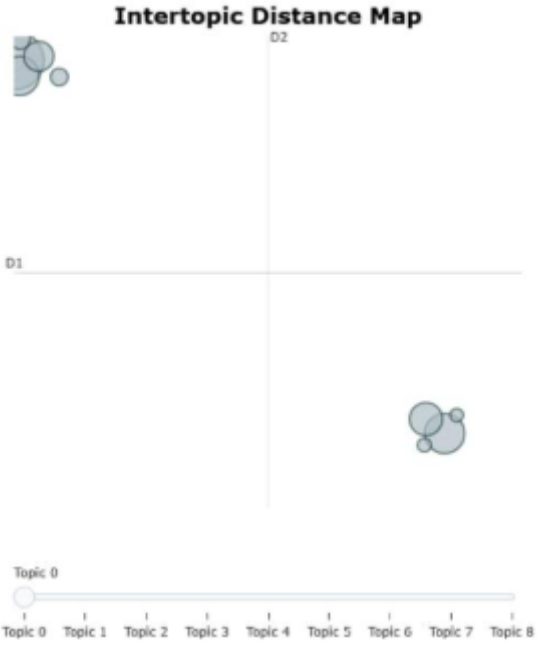
Pemodelan topik menggunakan metode BERTopic diawali dengan penggunaan data hasil preprocessing tanpa stemming, karena metode ini mempertahankan informasi morfologis dan semantik kata. Representasi teks dibentuk menggunakan CountVectorizer dengan parameter min_df=2 untuk mengabaikan kata yang jarang muncul dan ngram_range=(1,2) untuk menangkap konteks unigram dan bigram. Selanjutnya, setiap dokumen diubah menjadi embedding semantik menggunakan SentenceTransformer berbasis BERT, kemudian direduksi dimensinya menggunakan UMAP dengan parameter n_neighbors=15 dan n_components=10, sebelum dilakukan clustering menggunakan HDBSCAN.

Berbeda dengan LDA yang mengharuskan penentuan jumlah topik secara eksplisit di awal, BERTopic tidak menentukan jumlah topik secara langsung. Jumlah topik ditentukan secara otomatis berdasarkan hasil clustering menggunakan algoritma HDBSCAN yang bekerja berdasarkan kepadatan data. Penentuan model BERTopic yang optimal dilakukan melalui tuning parameter min_cluster_size pada HDBSCAN dengan rentang nilai 30–80. Parameter ini krusial dalam mengontrol jumlah topik karena BERTopic bekerja berdasarkan kepadatan data. Berdasarkan evaluasi menggunakan *coherence score*, hasil terbaik diperoleh pada min_cluster_size = 60 dengan skor 0,6064.

Table 3. Daftar Kata pada Topik BERTopic Paling Optimal

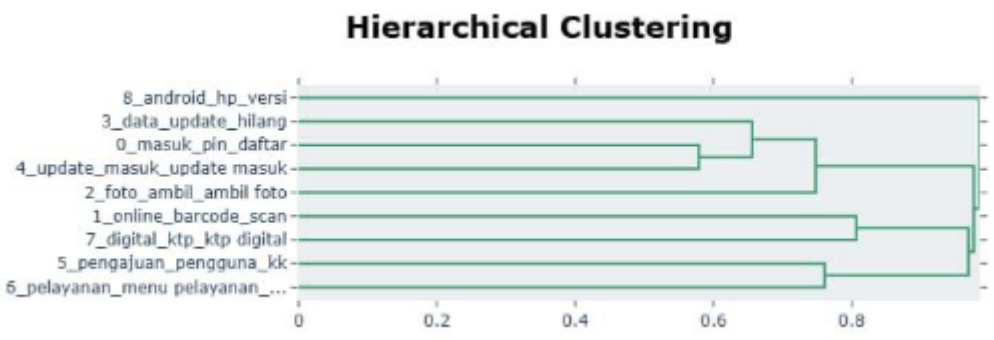
Topik	Daftar Kata
0	masuk (0.0494), pin (0.0486), daftar (0.0392), buka (0.0332), tolong (0.0280), koneksi (0.0261), akun (0.0230), bikin (0.0226), aplikasinya (0.0210), lupa (0.0199)
1	online (0.1282), barcode (0.1245), scan (0.0968), scan barcode (0.0720), dukcapil (0.0589), kantor (0.0569), petugas (0.0532), qr (0.0483), offline (0.0463), daftar (0.0445)
2	foto (0.1944), ambil (0.0433), ambil foto (0.0425), wajah (0.0406), frame (0.0393), ktp (0.0349), masuk (0.0343), selfie (0.0335), tombol (0.0305), foto wajah (0.0299)
3	data (0.1955), update (0.0531), hilang (0.0448), update data (0.0376), data hilang (0.0359), daftar (0.0316), ditemukan (0.0283), masuk (0.0264), data ditemukan (0.0253), valid (0.0245)
4	update (0.2251), masuk (0.1079), update masuk (0.0973), ulang (0.0474), akun (0.0423), habis update (0.0407), pin (0.0390), masuk update (0.0363), habis (0.0354), daftar (0.0324)
5	pengajuan (0.1263), pengguna (0.0476), kk (0.0451), pengurusan (0.0359), cetak kk (0.0327), pengajuan cetak (0.0308), cetak (0.0296), proses (0.0286), mohon (0.0284), akta (0.0278)
6	pelayanan (0.3328), menu pelayanan (0.1056), menu (0.1024), pelayanannya (0.0688), cetak (0.0547), buka (0.0502), kk (0.0409), tolong (0.0408), pelayanan buka (0.0388), lambat (0.0365)

7	digital (0.2073), ktp (0.1770), ktp digital (0.1517), bank (0.0508), digitalisasi (0.0404), fisik (0.0396), ktp fisik (0.0340), bikin (0.0323), dukcapil (0.0301), bikin ktp (0.0295)
8	android (0.3781), hp (0.1809), versi (0.1735), android versi (0.1494), hp android (0.0872), versi android (0.0801), support (0.0727), terbaru (0.0643), samsung (0.0580), android terbaru (0.0538)



Gambar 7. *Intertopic Distance Map* BERTopic

Meskipun BERTopic menghasilkan 9 topik dengan nilai *diversity* yang cukup tinggi (0,8222), hasil pada Tabel 3 yaitu distribusi kata kunci pada masing-masing topik menunjukkan adanya redundansi semantik antar topik. Indikasi redundansi tersebut diperkuat oleh visualisasi pada Gambar 7 yaitu Intertopic Distance Map yang menampilkan beberapa lingkaran topik saling bertumpang tindih (*overlap*) secara signifikan dan membentuk kelompok klaster yang sangat berdekatan, baik di kuadran kiri atas maupun kanan bawah. Kedekatan jarak antar-topik ini secara spasial menunjukkan bahwa topik-topik tersebut memiliki kemiripan konten yang tinggi dan tidak terpisah secara eksklusif. Fenomena ini merupakan dampak dari mekanisme clustering berbasis densitas pada HDBSCAN yang cenderung mengidentifikasi variasi ekspresi bahasa sebagai klaster yang terpisah, meskipun secara substantif membahas hal yang sama. Oleh karena itu, dilakukan proses penggabungan topik (topic merging) sebagai langkah pasca-pemrosesan[12]. Tahap ini bertujuan untuk mengeliminasi redundansi sehingga dihasilkan struktur topik yang lebih kohesif, solid, dan representatif terhadap pengalaman pengguna.



Gambar 8. *Intertopic Distance Map* BERTopic

Visualisasi dendrogram pada Gambar 7 menunjukkan bahwa Topik 0 dan 4 memiliki tingkat kedekatan tertinggi sebelum akhirnya menyatu dengan Topik 3 dalam satu kelompok. Pola serupa terlihat pada pasangan Topik 1 dan 7, serta Topik 5 dan 6 yang berbagi domain semantik dalam layanan administrasi kependudukan. Berdasarkan kedekatan struktur dan irisan kata kunci tersebut, dilakukan prosedur topic merging yang menghasilkan tiga kategori utama: kendala akses akun pasca-pembaruan (Topik 0, 3, 4), prosedur aktivasi melalui Dukcapil (Topik 1, 7), dan ulasan fitur layanan IKD (Topik 5, 6). Sebaliknya, Topik 2 (verifikasi wajah) dan Topik 8 (kompatibilitas Android) tetap dipertahankan sebagai entitas mandiri guna menjaga spesifisitas interpretasi, mengingat konteks keduanya yang unik.Tahap pasca-pemrosesan ini terbukti meningkatkan kualitas model secara signifikan. Skor koherensi (c_v) meningkat dari 0,6064 menjadi 0,6196, yang menandakan penguatan konsistensi semantik pada setiap topik. Selain itu, nilai topic diversity melonjak dari 0,8222 ke 0,92, mengonfirmasi bahwa penggabungan berhasil meminimalisir redundansi dan menghasilkan topik yang lebih distingtif serta representatif terhadap dataset. Melalui topic merging, jumlah topik disederhanakan menjadi lima kategori utama untuk meningkatkan keterwakilan data. Ekstraksi kata kunci dilakukan kembali guna memperoleh label yang lebih akurat dan jelas. Sebagaimana dirinci pada Tabel 4, topik-topik yang dihasilkan kini lebih distingtif, spesifik, dan tidak lagi menunjukkan redundansi yang signifikan.

Table 4. Daftar Kata pada Topik BERTopic Sebelum Merge

Topik	Daftar Kata	Interpretasi
0	masuk (0.0661), data (0.0585), update (0.0564), pin (0.0475), daftar (0.0453), buka (0.0322), ulang (0.0279), tolong (0.0277), akun (0.0272), aplikasinya (0.0237)	Login & Akses Akun
1	online (0.1360), barcode (0.1272), scan (0.1002), scan barcode (0.0712), dukcapil (0.0659), kantor (0.0603), petugas (0.0538), digital (0.0523), daftar (0.0506), qr (0.0503)	Scan Barcode ke Dukcapil
2	foto (0.2307), ambil (0.0488), ambil foto (0.0472), wajah (0.0467), frame (0.0439), masuk (0.0425), ktp (0.0419), selfie (0.0372), update (0.0348), tombol (0.0340)	Verifikasi Foto Wajah
3	pelayanan (0.1342), pengajuan (0.0956), kk (0.0508), menu (0.0468), cetak (0.0433), menu pelayanan (0.0380), pengguna (0.0356), cetak kk (0.0351), respon (0.0298), buka (0.0289)	Dokumen & Layanan Digital
4	android (0.4203), hp (0.2080), versi (0.1942), android versi (0.1621), hp android (0.0942), versi android (0.0865), support (0.0791), terbaru (0.0713), samsung (0.0625), android terbaru (0.0576)	Kompatibilitas Perangkat Android

3.7 Evaluasi Hasil Pemodelan Topik

Setelah proses pemodelan topik selesai dan jumlah topik optimal ditentukan berdasarkan nilai coherence score tertinggi, langkah selanjutnya adalah melakukan interpretasi topik. Tahap ini bertujuan untuk memahami makna yang terkandung di balik kumpulan kata yang dihasilkan oleh model LDA maupun BERTopic.

Table 5. Daftar Kata pada Topik BERTopic Sebelum Merge

Topic Modeling	Jumlah Topik	Coherence Score	Diversity
LDA	3	0.5083	0.9333
BERTopic	5	0.6196	0.92

Berdasarkan Tabel 5, terdapat perbedaan performa yang kontras antara metode LDA dan BERTopic. Keunggulan BERTopic terlihat dari skor koherensi sebesar 0,6196, melampaui LDA

yang hanya mencapai 0,5083. Hal ini membuktikan bahwa BERTopic menghasilkan topik yang lebih padu karena kemampuannya dalam menangkap makna kontekstual melalui arsitektur Transformer (Sentence-BERT), berbeda dengan LDA yang hanya berbasis pada distribusi probabilitas kata. Meskipun skor diversity BERTopic (0,92) sedikit di bawah LDA (0,9333), keduanya tetap berada pada kategori sangat tinggi. Dengan demikian, BERTopic dianggap sebagai model yang lebih optimal karena memberikan stabilitas semantik yang lebih baik tanpa mengurangi keragaman topik secara berarti.

4. KESIMPULAN DAN SARAN

Analisis komparatif menggunakan metrik Coherence Score dan Topic Diversity menunjukkan bahwa BERTopic memiliki performa yang lebih unggul dibandingkan Latent Dirichlet Allocation (LDA) dalam menghasilkan topik yang kohesif secara semantik dan representatif terhadap data ulasan aplikasi. Model LDA pada konfigurasi optimal menghasilkan tiga topik dengan skor koherensi 0,5083 dan diversity sebesar 0,9333. Walaupun skor keberagaman LDA lebih tinggi, topik yang dihasilkan bersifat generik dan kurang spesifik, seperti terlihat pada Topik 0 yang menggabungkan berbagai konteks permasalahan dalam satu kategori. Fenomena ini merupakan dampak dari ketergantungan LDA pada pendekatan probabilistik berbasis frekuensi kata, yang sering kali gagal menangkap nuansa semantik yang lebih kompleks [7]. Sebaliknya, BERTopic dengan parameter min_cluster_size = 60 menghasilkan lima topik pasca-penggabungan (merging) dengan skor koherensi 0,6196 dan keberagaman 0,92. Meskipun terdapat sedikit penurunan pada skor diversity, BERTopic mampu menghasilkan topik yang lebih granular dan spesifik melalui frasa teknis seperti "scan barcode" dan "verifikasi foto wajah" yang tidak terdeteksi oleh LDA. Keunggulan ini karena penggunaan transformer-based embeddings melalui Sentence-BERT, yang memungkinkan model memahami konteks kata secara lebih mendalam dibandingkan model tradisional[13]. Selain itu, BERTopic memiliki efektivitas lebih tinggi pada data ulasan yang pendek dan informal karena tidak memerlukan prosedur stemming yang agresif, sehingga tetap mempertahankan informasi kontekstual dari struktur kalimat asli[6]. Berdasarkan hasil penelitian, lima aspek utama yang berhasil diidentifikasi oleh BERTopic, yaitu “Login & Akses Akun”, “Scan Barcode ke Dukcapil”, “Verifikasi Foto Wajah”, “Dokumen & Layanan Digital”, serta “Kompatibilitas Perangkat Android”, dinilai lebih representatif dalam menggambarkan pengalaman dan permasalahan pengguna aplikasi IKD. Dengan demikian, BERTopic dapat disimpulkan sebagai metode yang lebih optimal dibandingkan LDA untuk pemodelan topik pada ulasan aplikasi IKD karena mampu menghasilkan topik yang lebih spesifik, mudah diinterpretasikan, dan memiliki kualitas semantik yang lebih baik.

5. DAFTAR RUJUKAN

[1] Irma Nurdiana and Khithoh Ayumi, “Implementasi Aplikasi Identitas Kependudukan Digital (IKD) Di Disdukcapil Kota Tanjungpinang,” *Harmoni Sos. J. Pengabdi. Dan Solidar. Masy.*, vol. 1, no. 2, pp. 50–58, Apr. 2024, doi: 10.62383/harmoni.v1i2.141.

[2] P. Hakiki, D. Satria, and A. A. Arifiyanti, “Prediksi Sentimen dan Pemodelan Topik dari Ulasan Aplikasi Identitas Kependudukan Digital,” *Jutisi J. Ilm. Tek. Inform. Dan Sist. Inf.*, vol. 14, no. 1, p. 760, Jul. 2025, doi: 10.35889/jutisi.v14i1.2777.

[3] Kementerian Dalam Negeri, “Peraturan Menteri Dalam Negeri Nomor 72 Tahun 2022 tentang Standar dan Spesifikasi Perangkat Keras, Perangkat Lunak, dan Blangko Kartu Tanda Penduduk Elektronik serta Penyelenggaraan Identitas Kependudukan Digital.” 2022. [Online]. Available: https://peraturan.bpk.go.id/Details/247759/permendagri-no-72-tahun-2022

[4] D. E. Sondakh, R. C. Maringka, F. P. Ayorbaba, J. S. C. B. T. Mangi, and S. R. Pungus, “Emotion Mining User Review of the BRImo Mobile Banking Application Using the Decision Tree Algorithm,” *J. Sisfokom Sist. Inf. Dan Komput.*, vol. 12, no. 3, pp. 350–355, Nov. 2023, doi: 10.32736/sisfokom.v12i3.1721.

[5] C. S. Ponay, "Topic Modeling on Customer Feedback from an Online Ticketing System using Latent Dirichlet Allocation and BERTopic," in *2022 2nd International Conference in Information and Computing Research (iCORE)*, Cebu, Philippines: IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/iCORE58172.2022.00020.
[6] M. D. Pratiwi and K. D. Tania, "Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMF," *J. Appl. Inform. Comput.*, vol. 9, no. 1, pp. 1–7, Jan. 2025, doi: 10.30871/jaic.v9i1.8782.
[7] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, p. 886498, May 2022, doi: 10.3389/fsoc.2022.886498.
[8] A. Hosseiny Marani and E. P. S. Baumer, "A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability," *ACM Comput. Surv.*, vol. 56, no. 5, pp. 1–32, May 2024, doi: 10.1145/3623269.
[9] R. Li, F. González-Pizarro, L. Xing, G. Murray, and G. Carenini, "Diversity-Aware Coherence Loss for Improving Neural Topic Models," 2023, *arXiv*. doi: 10.48550/ARXIV.2305.16199.
[10] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Appl. Sci.*, vol. 13, no. 2, p. 797, Jan. 2023, doi: 10.3390/app13020797.
[11] T. Gokcimen and B. Das, "Exploring climate change discourse on social media and blogs using a topic modeling analysis," *Heliyon*, vol. 10, no. 11, p. e32464, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32464.
[12] W. Janssens, M. Bogaert, and D. Van Den Poel, "A Comparative Analysis of Topic Reduction Techniques for BERTopic," *IEEE Access*, vol. 13, pp. 204087–204103, 2025, doi: 10.1109/ACCESS.2025.3638956.
[13] A. Riaz, O. Abdulkader, M. J. Ikram, and S. Jan, "Exploring topic modelling: a comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity," *Int. J. Electr. Comput. Eng. IJECE*, vol. 15, no. 2, p. 1933, Apr. 2025, doi: 10.11591/ijece.v15i2.pp1933-1948.