Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

Analisa Perbandingan Latent Semantic Indexing (LSI) dan Latent Dirichlet Allocation (LDA) untuk Topic Modelling Aplikasi Identitas Kependudukan Digital (IKD)

Nuri Cahyono[1,*], Narwanto Nurcahyo[2], Akmal Fauzan Restu Agung[2]

[1]Fakultas Ilmu Komputer, S1 Informatika, Universitas, Amikom Yogyakarta, Sleman, Indonesia
[2]Fakultas Ekonomi Sosial, S1 Kewirausahaan, Amikom Yogyakarta, Sleman, Indonesia
Email: [1,*]nuricahyono@amikom.ac.id, [2]narwanto.n@amikom.ac.id, [3]aafauzan@students.amikom.ac.id
Email Penulis Korespondensi: nuricahyono@amikom.ac.id
Submitted: **25/09/2024**; Accepted: **13/12/2024**; Published**: 18/12/2024**

Abstrak−Penelitian ini bertujuan untuk menganalisis dan membandingkan dua metode *Topic Modelling*, yaitu *Latent Semantic Indexing* (LSI) dan *Latent Dirichlet Allocation* (LDA), dalam memahami ulasan pengguna Aplikasi Identitas Kependudukan Digital (IKD) yang diperoleh dari Google Play Store. Masalah utama yang dihadapi adalah banyaknya ulasan pengguna dengan topik beragam yang sulit dikelompokkan secara manual, sehingga diperlukan metode otomatis untuk mengidentifikasi tema utama dalam data tersebut. Proses penelitian dimulai dengan scraping 5.000 ulasan terbaru, diikuti dengan prapemrosesan data *(Remove Punctuation, Lowercase, dan Tokenization)* serta vektorisasi menggunakan *Bag of Words* dan DOC2BOW. Selanjutnya, *Topic Modelling* dilakukan menggunakan metode LSI dan LDA, dan hasilnya dievaluasi menggunakan metrik *Coherence Score*. Hasil penelitian menunjukkan bahwa *Latent Dirichlet Allocation* (LDA) memiliki performa lebih baik dengan *Coherence Score* sebesar 0.4163, dibandingkan LSI yang hanya mencapai 0.3512, yang berarti *Latent Dirichlet Allocation* (LDA) lebih efektif dalam mengidentifikasi topik-topik yang tersembunyi dalam ulasan pengguna. *Latent Dirichlet Allocation* (LDA) merupakan metode yang lebih unggul untuk *Topic Modelling* ulasan aplikasi IKD dan dapat membantu pengembang memahami kebutuhan dan masalah pengguna, sehingga dapat meningkatkan kualitas layanan aplikasi.

Kata Kunci: Coherence Score; Identitas Kependudukan Digital (IKD); Latent Dirichlet Allocation (LDA); Latent Semantic Indexing (LSI); Topic Modeling

Abstract−This study aims to analyze and compare two topic modeling methods, Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA), in understanding user reviews of the Digital Population Identity (IKD) Application obtained from the Google Play Store. The main problem addressed is the large number of user reviews with diverse topics that are difficult to categorize manually, necessitating an automated method to identify the main themes in the data. The research process began with scraping 5,000 recent reviews, followed by data preprocessing (Remove Punctuation, Lowercase, and Tokenization) and vectorization using Bag of Words and DOC2BOW. Subsequently, topic modeling was performed using LSI and LDA, and the results were evaluated using the Coherence Score metric. The findings indicated that Latent Dirichlet Allocation (LDA) outperformed LSI, achieving a Coherence Score of 0.4163 compared to LSI's 0.3512, indicating that Latent Dirichlet Allocation (LDA) is more effective in identifying hidden topics within user reviews. Latent Dirichlet Allocation (LDA) is a superior method for topic modeling in IKD application reviews and can assist developers in understanding user needs and issues, thereby enhancing the application's service quality.

Keywords: Coherence Score; Identitas Kependudukan Digital (IKD); Latent Dirichlet Allocation (LDA); Latent Semantic Indexing (LSI); Topic Modeling

1. PENDAHULUAN

Penggunaan Aplikasi Identitas Kependudukan Digital (IKD) menunjukkan tren peningkatan yang signifikan, sejalan dengan inisiatif pemerintah untuk memaksimalkan layanan administrasi kependudukan melalui *platform* digital[1]. Menurut data yang dirilis oleh Kementerian Dalam Negeri pada 24 Mei 2024, jumlah individu yang telah menggunakan IKD mencapai 9.407.945 orang. Angka ini mencerminkan potensi yang besar dalam pemanfaatan teknologi digital untuk pengelolaan data kependudukan di Indonesia[2]. Situasi ini menunjukkan adanya kebutuhan mendesak akan sistem yang lebih efisien dalam pengelolaan informasi, termasuk dalam hal memahami umpan balik dan komentar dari pengguna terkait aplikasi IKD. Ulasan dari pengguna memiliki peranan yang sangat penting dalam menilai kinerja aplikasi serta memberikan informasi yang berharga untuk perbaikan dan pengembangan layanan yang lebih baik[3].

Proses pengelolaan dan analisis umpan balik dari pengguna IKD masih menghadapi sejumlah tantangan, terutama yang berkaitan dengan meningkatnya volume data yang semakin kompleks dan bervariasi[4]. Kesulitan dalam mengekstrak informasi penting dari data ulasan yang bersifat tidak terstruktur sering kali mengakibatkan proses analisis menjadi kurang efisien dan memakan waktu yang cukup lama. Selain itu, beragamnya topik yang muncul dalam ulasan pengguna menyulitkan para pengembang aplikasi untuk memahami permasalahan utama dan kebutuhan pengguna secara menyeluruh. Akibatnya, banyak informasi penting yang seharusnya dapat dimanfaatkan untuk meningkatkan kualitas layanan IKD tidak teridentifikasi dengan baik.

Untuk mengatasi permasalahan tersebut, teknologi *Topic Modelling* dapat menjadi solusi efektif dalam mengidentifikasi tema utama dari data teks ulasan pengguna. *Topic Modelling* memungkinkan analisis data teks secara otomatis sehingga ulasan pengguna dapat dikelompokkan berdasarkan topik yang relevan. Hal ini memberi peluang bagi pengembang aplikasi untuk memahami kebutuhan dan permasalahan pengguna dengan lebih efisien. Di antara

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

metode *Topic Modelling* yang tersedia, dua pendekatan yang sering digunakan adalah *Latent Semantic Indexing* (LSI) dan *Latent Dirichlet Allocation* (LDA). Kedua metode ini efektif dalam mengekstraksi informasi *late*n dari kumpulan data teks yang besar.

Latent Semantic Indexing (LSI) adalah metode yang menggunakan teknik dekomposisi nilai singular *(Singular Value Decomposition)* untuk mengidentifikasi hubungan laten antara kata dan dokumen dalam ruang dimensi yang lebih rendah[5]. Metode ini dapat mengungkap pola dan hubungan semantik antar kata yang sering muncul bersama dalam ulasan pengguna[6]. Sebaliknya, *Latent Dirichlet Allocation* (LDA) adalah pendekatan probabilistik yang memodelkan ulasan pengguna sebagai campuran dari beberapa topik, dengan setiap topik direpresentasikan oleh distribusi sejumlah kata[7]. LDA memiliki kemampuan untuk menggambarkan distribusi kata dalam topik secara lebih fleksibel, sehingga memungkinkan identifikasi topik-topik tersembunyi dalam ulasan pengguna secara lebih mendetail[8].

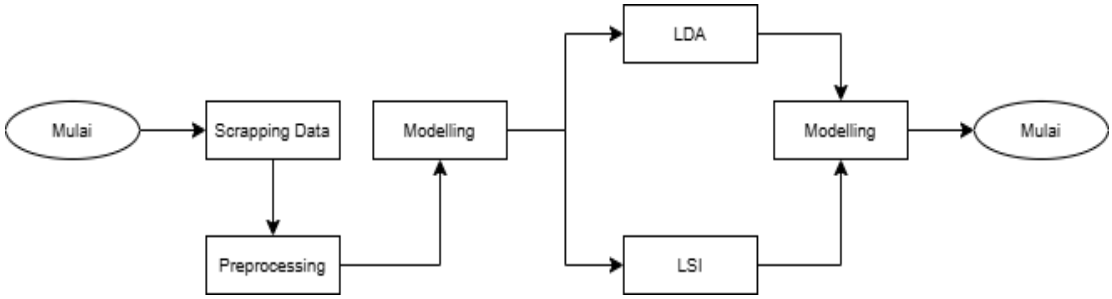
Hasil dari penelitian sebelumnya menunjukkan bahwa metode *Latent Dirichlet Allocation* (LDA) telah banyak digunakan dalam berbagai konteks pemodelan topik, tetapi umumnya masih terbatas pada penerapan satu metode tanpa perbandingan. Astuti dan Cahyono menerapkan LDA untuk menganalisis berita di portal *online* detik.com, mengidentifikasi tiga topik utama yaitu bencana alam, isu politik, dan Piala Dunia[9]. Sementara itu, Khadijah dan Cahyono menggunakan LDA untuk menganalisis topik pariwisata di Yogyakarta berdasarkan data Twitter dan menemukan tiga topik terkait aktivitas sehari-hari, penawaran pariwisata, dan informasi infrastruktur[7]. Di bidang pendidikan, Rosales et al. menggunakan LDA untuk mengidentifikasi 10 topik terkait metodologi dan strategi pengajaran inovatif di universitas[10]. Meskipun hasilnya efektif, penelitian-penelitian ini tidak melakukan perbandingan dengan metode lain, sehingga terbatas dalam mengeksplorasi efektivitas pemodelan topik dari berbagai pendekatan.

Penelitian lainnya oleh Kannitha et al. menggunakan LDA untuk mengidentifikasi keluhan pelanggan layanan internet melalui data dari Twitter dan menemukan bahwa LDA berhasil mengidentifikasi topik keluhan secara akurat [8]. Singgalen juga menerapkan pendekatan LDA dalam mengoptimalkan pemasaran destinasi pariwisata di Indonesia dan mengidentifikasi topik terkait fasilitas, daya tarik wisata, dan aksesibilitas[11]. Meskipun LDA terbukti efektif dalam berbagai studi, semua penelitian tersebut hanya menggunakan LDA tanpa mengevaluasi performa metode lain seperti *Latent Semantic Indexing* (LSI), sehingga potensi efektivitas metode alternatif belum dieksplorasi.

Meskipun LSI dan LDA telah diterapkan dalam berbagai penelitian mengenai *Topic Modelling*, sebagian besar studi cenderung hanya berfokus pada satu metode tanpa melakukan analisis perbandingan antara keduanya, terutama dalam konteks ulasan pengguna Aplikasi IKD. Penelitian ini bertujuan untuk mengatasi kekurangan tersebut dengan melakukan analisis menyeluruh terhadap kinerja LSI dan LDA dalam *Topic Modelling* ulasan pengguna IKD. Analisis perbandingan antara LSI dan LDA dalam konteks ulasan pengguna Aplikasi IKD dapat memberikan wawasan berharga tentang kelebihan dan kekurangan masing-masing metode. Dengan melakukan evaluasi menyeluruh terhadap kedua teknik *Topic Modelling* ini, penelitian ini diharapkan dapat mengungkapkan pola-pola tersembunyi dan tren yang mungkin tidak terdeteksi jika hanya menggunakan satu metode saja. Hasil dari analisis komparatif ini dapat membantu para peneliti dalam memilih pendekatan yang paling efektif untuk menganalisis umpan balik pengguna dan membantu memilih metode *Topic Modelling* yang paling sesuai untuk analisis ulasan pengguna Aplikasi IKD di masa depan.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk melakukan analisis perbandingan antara metode *Latent Semantic Indexing* (LSI) dan *Latent Dirichlet Allocation* (LDA) dalam *Topic Modelling* ulasan pengguna Aplikasi Identitas Kependudukan Digital (IKD) yang diperoleh dari Google Play Store. Penelitian ini terdiri dari beberapa tahap, yaitu pengambilan data ulasan, prapemrosesan data, representasi data dengan teknik *Bag of Words* (BoW), *Topic Modelling* menggunakan LSI dan LDA, evaluasi model, dan visualisasi topik yang dihasilkan. Setiap tahap dalam metodologi penelitian dijelaskan secara rinci pada Gambar 1.



Gambar 1. Tahapan Penelitian

Tahap pertama adalah pengumpulan data ulasan pengguna aplikasi IKD melalui proses *scraping* dari Google Play Store. *Scraping* data dilakukan dengan menggunakan perangkat lunak otomatis yang dapat mengambil data

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

ulasan pengguna secara langsung dari *platform* Google Play Store[12]. Dalam penelitian ini, ulasan pengguna dikumpulkan secara lengkap, termasuk judul ulasan, konten ulasan, skor penilaian, tanggal ulasan, dan informasi lainnya yang relevan. Proses ini sangat penting karena data ulasan pengguna akan menjadi sumber utama dalam *Topic Modelling*[13]. Data ini kemudian disimpan dalam format terstruktur seperti file CSV atau *database*, sehingga memudahkan pengolahan dan analisis lebih lanjut. Teknik *scraping* memungkinkan peneliti untuk mengumpulkan sejumlah besar data ulasan pengguna dalam waktu singkat, yang diperlukan untuk menghasilkan *Topic Modelling* yang komprehensif dan representatif.

Setelah data ulasan berhasil diperoleh, tahap berikutnya adalah melakukan prapemrosesan data teks. Prapemrosesan ini bertujuan untuk membersihkan dan menyiapkan data sehingga dapat diolah oleh model *Topic Modelling*. Tahapan prapemrosesan meliputi langkah-langkah berikut:

a. *Remove Punctuation:* Menghapus tanda baca seperti titik, koma, tanda seru, tanda tanya, dan karakter khusus lainnya yang tidak memiliki nilai informasi dalam analisis teks. Hal ini dilakukan untuk memastikan bahwa data hanya terdiri dari kata-kata yang relevan.
b. *Lowercase:* Mengubah semua huruf dalam teks menjadi huruf kecil untuk menghindari perbedaan format yang disebabkan oleh huruf kapital. Misalnya, kata "Aplikasi" dan "aplikasi" akan dianggap sebagai kata yang sama setelah diubah menjadi huruf kecil.
c. *Tokenization:* Memecah setiap kalimat dalam ulasan menjadi kata-kata individu atau token. Proses ini membantu dalam mengidentifikasi kata-kata kunci dalam ulasan pengguna yang akan digunakan dalam *Topic Modelling*.

Prapemrosesan ini memastikan bahwa data teks yang akan digunakan dalam tahap pemodelan bebas dari elemen-elemen yang dapat mengganggu analisis, sehingga menghasilkan model yang lebih akurat dan efisien. Langkah selanjutnya adalah membangun representasi data teks menggunakan teknik *Bag of Words* (BoW). BoW adalah pendekatan yang mengubah data teks menjadi vektor numerik berdasarkan frekuensi kemunculan kata dalam dokumen[14]. Setiap ulasan pengguna direpresentasikan sebagai vektor yang menunjukkan berapa kali setiap kata muncul dalam ulasan tersebut[15]. Representasi ini memungkinkan model untuk memahami data teks sebagai sekumpulan fitur numerik yang dapat diproses oleh algoritma pembelajaran mesin. Dengan menggunakan BoW, model dapat mengenali pola dan frekuensi kata yang muncul dalam ulasan pengguna, yang akan digunakan untuk mengidentifikasi topik utama dalam data teks[16].

Tahapan selanjutnya dalam metodologi penelitian ini adalah melakukan visualisasi data topik yang dihasilkan oleh model LDA dan LSI. Visualisasi topik dilakukan untuk memberikan gambaran yang lebih jelas dan intuitif mengenai topik-topik yang teridentifikasi dari data ulasan. Teknik visualisasi, seperti *word cloud* atau grafik distribusi topik, akan digunakan untuk menampilkan kata-kata kunci yang paling sering muncul dalam setiap topik serta distribusi topik dalam ulasan pengguna. Visualisasi ini membantu peneliti dan pengembang aplikasi dalam memahami tema utama yang terdapat dalam ulasan pengguna dan bagaimana topik tersebut tersebar di seluruh data ulasan.

Setelah model LDA dan LSI selesai dibentuk, tahap terakhir adalah melakukan evaluasi model menggunakan *metrik Coherence Score* [17]. *Coherence Score* adalah metrik yang mengukur seberapa koheren atau konsisten topik-topik yang dihasilkan oleh model[18]. Koherensi topik menggambarkan sejauh mana kata-kata dalam suatu topik saling terkait dan relevan satu sama lain[19]. Semakin tinggi nilai *Coherence Score*, semakin baik kualitas topik yang dihasilkan oleh model[20]. Dalam penelitian ini, nilai *Coherence Score* dari model LDA dan LSI akan dibandingkan untuk menentukan metode *Topic Modelling* mana yang paling efektif dalam analisis ulasan pengguna aplikasi IKD.

3. HASIL DAN PEMBAHASAN

3.1 *Scrapping Data*

Penelitian ini dimulai dengan proses scraping data ulasan pengguna Aplikasi Identitas Kependudukan Digital (IKD) dari Google Play Store. *Scraping* dilakukan untuk memperoleh data ulasan terbaru hingga bulan Agustus 2024. Proses pengambilan data ini menggunakan teknik *scraping* berbasis program otomatis yang memungkinkan peneliti mengumpulkan informasi yang relevan dari aplikasi tersebut secara efisien. Dari hasil scraping, diperoleh sebanyak 5.000 ulasan pengguna yang dipilih berdasarkan ulasan terbaru hingga batas waktu yang ditentukan. Pengambilan data dalam jumlah ini bertujuan untuk memastikan bahwa analisis yang dilakukan mencerminkan kondisi terkini dan mencakup berbagai pandangan pengguna tentang aplikasi IKD.

Parameter data yang diambil selama proses *scraping* meliputi berbagai aspek penting, yaitu: *reviewId*, *userName*, *userImage*, *content*, *score*, *thumbsUpCount*, *reviewCreatedVersion*, *at*, *replyContent*, *repliedAt*, dan *appVersion*. Variabel-variabel ini memberikan informasi yang cukup komprehensif untuk analisis lebih lanjut. Misalnya, content berisi teks ulasan yang menjadi fokus utama dalam *Topic Modelling*, sedangkan *score* dan *thumbsUpCount* memberikan gambaran tentang tingkat kepuasan pengguna dan tingkat relevansi ulasan tersebut menurut pengguna lain. Parameter *reviewCreatedVersion* dan *appVersion* memungkinkan peneliti untuk mengidentifikasi apakah ulasan yang diberikan pengguna terkait dengan versi aplikasi tertentu, yang dapat menjadi indikasi masalah atau fitur yang spesifik pada versi tersebut.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

3.2 Preprocessing

Proses pembersihan data *(data cleaning)* dan prapemrosesan data *(preprocessing)* yang bertujuan untuk mempersiapkan data ulasan pengguna Aplikasi Identitas Kependudukan Digital (IKD) agar siap untuk dianalisis lebih lanjut. Pada tahap data cleaning, data hasil scraping yang diperoleh dari Google Play Store difokuskan hanya pada *parameter content*, yaitu bagian yang berisi teks ulasan pengguna terhadap aplikasi IKD. Hal ini dilakukan karena parameter content mengandung informasi yang relevan dan menjadi sumber utama dalam analisis *Topic Modelling*. Data ulasan lainnya yang tidak terkait langsung dengan analisis teks, seperti *reviewId*, *userName*, dan *score*, diabaikan dalam tahap ini. Dengan hanya mengambil parameter content, peneliti memastikan bahwa data yang digunakan benar-benar berfokus pada isi ulasan yang menjadi target penelitian.

Tabel 1. *Remove Punctuation*

Sebelum	Sesudah
"Aplikasi ini sangat bagus! Membantu dalam pengurusan dokumen."	"Aplikasi ini sangat bagus Membantu dalam pengurusan dokumen"
"Gagal login, selalu muncul pesan 'error' setiap kali mencoba."	"Gagal login selalu muncul pesan error setiap kali mencoba"
"Versi terbaru sering crash, mohon segera diperbaiki."	"Versi terbaru sering crash mohon segera diperbaiki"
"Sudah di-update, tapi tetap saja tidak bisa digunakan!"	"Sudah diupdate tapi tetap saja tidak bisa digunakan"
"Tolong tambahkan fitur pencarian! Akan sangat membantu."	"Tolong tambahkan fitur pencarian Akan sangat membantu"

Pada Tabel 1 semua tanda baca seperti tanda seru, titik, koma, tanda kutip, dan karakter khusus lainnya dihapus dari data ulasan pengguna. Penghapusan tanda baca ini penting untuk memastikan bahwa data ulasan terdiri dari kata-kata yang relevan dan bersih, sehingga memudahkan proses analisis berikutnya. Hasil dari tahap ini adalah data teks yang lebih sederhana dan bebas dari elemen yang tidak diperlukan.

Tabel 2. *Lowercase*

Sebelum	Sesudah
"Aplikasi ini sangat bagus Membantu dalam pengurusan dokumen"	"aplikasi ini sangat bagus membantu dalam pengurusan dokumen"
"Gagal login selalu muncul pesan error setiap kali mencoba"	"gagal login selalu muncul pesan error setiap kali mencoba"
"Versi terbaru sering crash mohon segera diperbaiki"	"versi terbaru sering crash mohon segera diperbaiki"
"Sudah diupdate tapi tetap saja tidak bisa digunakan"	"sudah diupdate tapi tetap saja tidak bisa digunakan"
"Tolong tambahkan fitur pencarian Akan sangat membantu"	"tolong tambahkan fitur pencarian akan sangat membantu"

Tabel 3. *Tokenization*

Sebelum	0Sesudah
"Aplikasi ini sangat bagus Membantu dalam pengurusan dokumen"	"aplikasi ini sangat bagus membantu dalam pengurusan dokumen"
"Gagal login selalu muncul pesan error setiap kali mencoba"	"gagal login selalu muncul pesan error setiap kali mencoba"
"Versi terbaru sering crash mohon segera diperbaiki"	"versi terbaru sering crash mohon segera diperbaiki"
"Sudah diupdate tapi tetap saja tidak bisa digunakan"	"sudah diupdate tapi tetap saja tidak bisa digunakan"
"Tolong tambahkan fitur pencarian Akan sangat membantu"	"tolong tambahkan fitur pencarian akan sangat membantu"

Pada Tabel 2 semua huruf dalam data ulasan diubah menjadi huruf kecil. Proses ini dilakukan untuk menjaga konsistensi dan menghindari masalah perbedaan format yang dapat terjadi jika kata yang sama ditulis dengan kombinasi huruf kapital dan kecil. Dengan menerapkan *lowercasing*, kata-kata seperti "Aplikasi" dan "aplikasi" diperlakukan sebagai entitas yang sama, sehingga mengurangi redundansi dalam data dan memastikan keakuratan analisis. Tahap terakhir adalah *Tokenization*, di mana teks ulasan yang telah melalui proses *Remove Punctuation* dan *Lowercase* dipecah menjadi kata-kata individu atau token. Proses *Tokenization* ini memisahkan setiap kalimat menjadi daftar kata, sehingga model *Topic Modelling* dapat mengidentifikasi dan menganalisis kata-kata yang sering muncul

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

dalam ulasan pengguna. Hasil dari tahap ini adalah data teks dalam bentuk kata-kata terpisah yang siap digunakan untuk analisis lebih lanjut pada Tabel 3.

Proses vektorisasi yang bertujuan untuk mengubah data teks ulasan aplikasi IKD menjadi format numerik yang dapat diproses oleh model *Topic Modelling*. Proses ini diawali dengan pembuatan *dictionary* dan *corpus*. *Dictionary* dibuat dengan mengidentifikasi semua kata unik yang terdapat dalam data ulasan hasil prapemrosesan sebelumnya (token yang telah dihasilkan melalui proses *tokenization*). *Dictionary* ini berfungsi sebagai referensi yang menghubungkan setiap kata dengan indeks tertentu, sehingga setiap kata dalam data teks memiliki representasi numerik. Sebagai contoh, kata-kata seperti "aplikasi," "bagus," dan "membantu" masing-masing akan memiliki indeks unik dalam *dictionary*, menciptakan struktur data yang memungkinkan model mengenali kata-kata tersebut dengan lebih mudah

Setelah *dictionary* terbentuk, langkah selanjutnya adalah mengonversi data teks ulasan menjadi bentuk numerik berdasarkan metode *Bag of Words* (BoW). Proses ini dilakukan dengan menghitung frekuensi kemunculan setiap kata dalam ulasan dan kemudian merepresentasikannya sebagai vektor numerik. Sebagai contoh, jika dalam suatu ulasan pengguna terdapat kata "aplikasi" sebanyak dua kali dan kata "bagus" satu kali, maka vektor ulasan tersebut akan memiliki nilai yang mencerminkan frekuensi kemunculan kata-kata tersebut sesuai dengan indeks yang ada dalam *dictionary*. Vektorisasi berbasis BoW ini menghasilkan representasi numerik yang menunjukkan seberapa sering setiap kata muncul dalam ulasan pengguna, memungkinkan model untuk memahami pola dan hubungan antar kata.

Proses terakhir dalam tahap vektorisasi adalah mengubah data teks ulasan ke dalam bentuk numerik menggunakan metode *DOC2BOW* (*Document to Bag of Words*). Metode *DOC2BOW* mengambil setiap ulasan pengguna yang telah diproses dan mengonversinya menjadi daftar pasangan yang berisi indeks kata dan frekuensi kemunculannya. Sebagai contoh, jika suatu ulasan berisi kata "aplikasi" dengan indeks 1 yang muncul dua kali dan kata "bagus" dengan indeks 2 yang muncul satu kali, maka hasil konversi menggunakan *DOC2BOW* akan menjadi [(1, 2), (2, 1)]. Proses ini menghasilkan *corpus* dalam bentuk numerik yang siap untuk digunakan dalam *Topic Modelling*, baik dengan metode LDA maupun LSI. Dengan melakukan vektorisasi, penelitian ini telah berhasil mengubah data teks menjadi representasi numerik yang dapat dianalisis secara matematis oleh model, memungkinkan identifikasi topik-topik utama dalam ulasan pengguna aplikasi IKD secara lebih efektif dan efisien.

Tabel 4. Representasi DOC2BOW

Ulasan	Representasi
"aplikasi ini sangat bagus membantu pengurusan dokumen"	[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]
"aplikasi sangat membantu"	[(0, 1), (2, 1)]

Pada Tabel 4 setiap ulasan dikonversi menjadi pasangan nilai (indeks kata, frekuensi kemunculan). Sebagai contoh, ulasan pertama menghasilkan pasangan [(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)], yang menunjukkan bahwa kata dengan indeks 0 ("aplikasi") muncul satu kali, indeks 1 ("bagus") muncul satu kali, dan seterusnya. Untuk ulasan kedua, hanya kata "aplikasi" dan "membantu" yang muncul, menghasilkan pasangan [(0, 1), (2, 1)].

3.3 Modelling LDA

```
(0, '0.067*"aplikasi" + 0.023*"gak" + 0.019*"bikin" + 0.017*"perintah" + 0.014*"aja"')
(1, '0.036*"aplikasi" + 0.023*"yg" + 0.019*"gak" + 0.011*"g" + 0.010*"hp"')
(2, '0.062*"buka" + 0.060*"aplikasi" + 0.032*"koneksi" + 0.029*"nya" + 0.027*"salah"')
(3, '0.030*"pin" + 0.029*"masuk" + 0.025*"kali" + 0.022*"hp" + 0.021*"ulang"')
(4, '0.041*"scan" + 0.037*"dukcapil" + 0.036*"barcode" + 0.035*"aplikasi" + 0.025*"daftar"')
(5, '0.044*"data" + 0.039*"aman" + 0.017*"pribadi" + 0.014*"vaksin" + 0.014*"log"')
(6, '0.042*"aplikasi" + 0.030*"ktp" + 0.025*"dokumen" + 0.025*"mudah" + 0.019*"digital"')
(7, '0.051*"android" + 0.031*"hp" + 0.023*"versi" + 0.020*"aplikasi" + 0.014*"yg"')
(8, '0.091*"ktp" + 0.039*"digital" + 0.019*"nya" + 0.018*"fisik" + 0.015*"bikin"')
(9, '0.017*"ngak" + 0.016*"yg" + 0.009*"buang2" + 0.008*"ga" + 0.007*"close"')
```

Gambar 2. Penerapan LDA

Berdasarkan Gambar 2 hasil modelling LDA *(Latent Dirichlet Allocation)* pada data ulasan aplikasi Identitas Kependudukan Digital (IKD), terlihat bahwa pemodelan ini menghasilkan 10 topik utama, dengan masing-masing topik memiliki 5 kata dengan bobot tertinggi yang menggambarkan karakteristik dari setiap topik. Misalnya, Topik 0 didominasi oleh kata-kata seperti "aplikasi," "gak," "bikin," "perintah," dan "aja," yang mungkin menunjukkan adanya keluhan pengguna tentang kesulitan dalam menggunakan aplikasi atau fitur tertentu.

Pada Topik 1, kata-kata yang muncul seperti "aplikasi," "yg," "gak," dan "hp" mengindikasikan bahwa pengguna sering menyebutkan masalah yang berhubungan dengan perangkat (handphone) dan kinerja aplikasi. Topik 2 menonjol dengan kata-kata seperti "buka," "aplikasi," "koneksi," "nya," dan "salah," yang menunjukkan bahwa pengguna sering menghadapi masalah saat membuka aplikasi atau menghubungkannya, mungkin terkait dengan kesalahan koneksi atau akses.

Topik lainnya, seperti Topik 4, dengan kata-kata "scan," "dukcapil," "barcode," "aplikasi," dan "daftar," mengindikasikan masalah terkait proses pendaftaran atau verifikasi identitas melalui sistem yang mungkin

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

memerlukan pemindaian atau interaksi dengan data dari Dinas Kependudukan dan Catatan Sipil (Dukcapil). Sedangkan pada Topik 5, kata-kata "data," "aman," "pribadi," "vaksin," dan "log" menunjukkan adanya kekhawatiran atau diskusi terkait keamanan data pribadi pengguna.

Hasil pemodelan LDA ini memberikan gambaran tentang berbagai isu dan topik yang sering diungkapkan oleh pengguna dalam ulasan mereka mengenai aplikasi IKD. Topik-topik ini bisa menjadi acuan bagi pengembang aplikasi untuk memahami masalah dan kebutuhan pengguna, sehingga dapat memperbaiki dan meningkatkan fitur serta layanan aplikasi sesuai dengan masukan yang telah diperoleh. Analisis ini menunjukkan bahwa LDA mampu mengidentifikasi tema-tema utama yang tersembunyi dalam data teks ulasan pengguna, yang sangat bermanfaat untuk pengembangan dan pengelolaan aplikasi di masa mendatang.

Tabel 5. Probabilitas Topik LDA

konten	Topik	Probabilitas
gk	8	0.01288
aplikasi	2	0.060251
digital	8	0.038988
koneksi	2	0.031787
hp	1	0.009604
online	4	0.01955
mudah	1	0.007314
sangat	0	0.013542
dukcapil	4	0.03675
aplikasi	0	0.066902

Berdasarkan hasil analisis LDA *(Latent Dirichlet Allocation)* pada tabel 5, terlihat bahwa beberapa kata memiliki probabilitas yang berbeda dalam merepresentasikan topik tertentu dari ulasan pengguna aplikasi IKD. Misalnya, kata "aplikasi" muncul pada topik 0 dengan probabilitas tertinggi sebesar 0,066902, menunjukkan bahwa kata ini sangat relevan dan sering muncul dalam topik tersebut. Hal ini mengindikasikan bahwa topik 0 sangat berhubungan dengan diskusi atau keluhan pengguna mengenai aplikasi secara umum. Kata "dukcapil" juga memiliki probabilitas yang cukup tinggi (0,036750) dalam topik 4, yang kemungkinan menunjukkan pembahasan terkait layanan Dinas Kependudukan dan Catatan Sipil yang terintegrasi dengan aplikasi IKD. Kata "koneksi" muncul dalam topik 2 dengan probabilitas 0,031787, mengindikasikan bahwa masalah koneksi sering disebutkan dalam konteks topik tersebut, yang mungkin merujuk pada kendala teknis yang dihadapi pengguna.

Kata "digital" memiliki probabilitas 0,038988 dalam topik 8, yang menunjukkan bahwa topik ini terkait dengan aspek digital dari aplikasi IKD. Sedangkan kata "mudah" muncul dalam dua topik, yaitu topik 1 dengan probabilitas 0,007314 dan topik 5 dengan probabilitas 0,013542, menunjukkan bahwa kemudahan penggunaan menjadi perhatian pengguna di beberapa konteks yang berbeda.

Probabilitas kata-kata ini membantu dalam memahami seberapa kuat keterkaitan masing-masing kata dengan topik tertentu, memberikan gambaran yang lebih jelas tentang aspek-aspek yang menjadi fokus perhatian atau permasalahan pengguna terkait aplikasi IKD. Analisis ini dapat membantu pengembang aplikasi dalam mengidentifikasi area yang perlu diperbaiki atau dikembangkan lebih lanjut berdasarkan topik yang diungkapkan oleh pengguna melalui ulasan mereka.

Gambar 3. *Wordcloud* pemodelan LDA

Berdasarkan hasil visualisasi *word cloud* dari pemodelan LDA *(Latent Dirichlet Allocation)* pada Gambar 3 terlihat bahwa topik-topik yang dihasilkan memiliki kata-kata kunci yang cukup jelas dan relevan dalam menggambarkan isu-isu utama yang sering dibahas oleh pengguna aplikasi Identitas Kependudukan Digital (IKD).

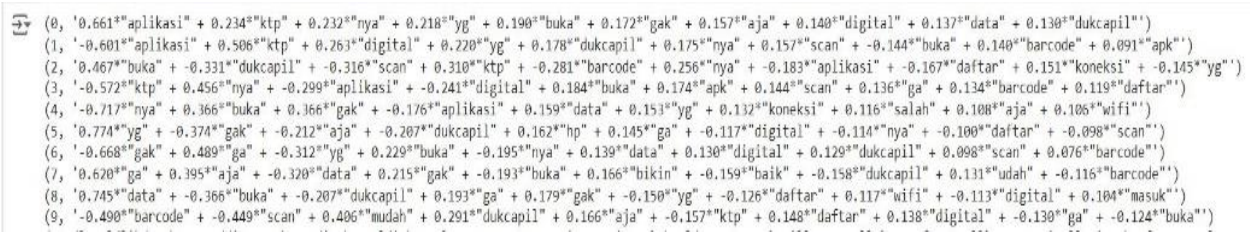
Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

Kata-kata seperti "aplikasi," "data," "dukcapil," "ktp," "scan," dan "masuk" muncul dengan ukuran yang lebih besar dalam beberapa topik, menandakan bahwa kata-kata ini sering muncul dalam ulasan pengguna dan memiliki keterkaitan yang kuat dengan pengalaman mereka saat menggunakan aplikasi IKD. Sebagai contoh, kata "aplikasi" muncul di hampir setiap topik, menunjukkan bahwa banyak pengguna memberikan ulasan mengenai pengalaman mereka terkait fitur, kinerja, atau kendala saat menggunakan aplikasi tersebut.

Topik lain menunjukkan kata-kata seperti "data," "dukcapil," dan "scan," yang mengindikasikan bahwa banyak pengguna membahas tentang proses verifikasi data, integrasi dengan layanan Dukcapil, dan penggunaan fitur scan dalam aplikasi IKD. Kata "aman" juga muncul sebagai kata kunci yang cukup signifikan, menandakan bahwa ada kekhawatiran atau perhatian pengguna terkait keamanan data pribadi mereka saat menggunakan aplikasi IKD. Selain itu, kata "masuk," "pin," dan "ulang" menunjukkan adanya masalah yang sering dihadapi oleh pengguna terkait proses login atau akses ke aplikasi.

Hasil *word cloud* dari LDA ini memberikan gambaran yang lebih terstruktur dan jelas mengenai topik-topik utama yang diangkat oleh pengguna dalam ulasan mereka. Berbeda dengan LSI, LDA mampu menangkap hubungan semantik yang lebih kuat antar kata-kata dalam topik, menghasilkan tema yang lebih koheren dan relevan. Visualisasi ini sangat membantu dalam memahami kebutuhan, masalah, dan persepsi pengguna, serta memberikan wawasan bagi pengembang aplikasi untuk melakukan perbaikan atau pengembangan fitur yang lebih sesuai dengan kebutuhan pengguna aplikasi IKD.

3.4 Modelling LSI



Gambar 4. Penerapan LSI

Berdasarkan hasil *modelling Latent Semantic Indexing* (LSI) pada Gambar 4 analisis *Topic Modelling* menghasilkan 10 topik utama dengan 5 kata yang paling signifikan pada masing-masing topik. Berbeda dengan LDA yang berbasis probabilistik, LSI menggunakan pendekatan linear aljabar untuk menemukan hubungan semantik antar kata dalam ulasan pengguna aplikasi IKD.

Pada Topik 0, kata-kata dengan bobot tertinggi adalah "aplikasi," "ktp," "nya," "yg," dan "buka." Hal ini menunjukkan bahwa topik ini kemungkinan besar berkaitan dengan pengalaman umum pengguna dalam menggunakan aplikasi, terutama dalam mengakses KTP digital. Topik ini dapat mencerminkan masalah teknis atau pengalaman penggunaan yang berkaitan dengan pengoperasian aplikasi tersebut.

Topik 1 menampilkan kata-kata seperti "aplikasi," "ktp," "digital," "yg," dan "dukcapil," yang menunjukkan adanya diskusi atau keluhan terkait integrasi data antara aplikasi IKD dan layanan Dinas Kependudukan dan Catatan Sipil (Dukcapil). Kata "digital" menunjukkan fokus pada aspek digitalisasi, mungkin terkait dengan proses verifikasi KTP atau data pengguna. Topik 2 menonjolkan kata-kata "buka," "dukcapil," "scan," "ktp," dan "barcode," yang mengindikasikan bahwa pengguna mungkin menghadapi masalah saat mencoba membuka atau memindai barcode untuk verifikasi data kependudukan melalui aplikasi. Topik ini menunjukkan adanya potensi masalah teknis dalam proses verifikasi identitas. Topik 5 menonjol dengan kata-kata seperti "yg," "gak," "aja," "ga," dan "digital," yang dapat menunjukkan ulasan yang mungkin berisi keluhan atau penilaian negatif dari pengguna terkait aspek tertentu dari aplikasi, terutama yang berkaitan dengan proses digital atau fitur yang dianggap tidak bekerja dengan baik.

Hasil pemodelan LSI ini memberikan gambaran yang lebih jelas tentang berbagai isu dan pengalaman pengguna yang terkait dengan aplikasi IKD. Analisis LSI mampu menyoroti hubungan semantik antar kata-kata dalam ulasan, membantu dalam memahami topik utama yang menjadi perhatian pengguna. Hal ini penting untuk mengidentifikasi area yang memerlukan perbaikan dan peningkatan dalam aplikasi IKD berdasarkan masukan pengguna.

Tabel 6. Probabilitas Topik LSI

konten	Topik	Probabilitas
nya	1	0.175353
baik	7	-0.158602
dukcapil	2	-0.330946
buka	0	0.190382
mudah	9	0.405948
wifi	8	0.117265
masuk	8	0.104064
dukcapil	5	-0.207045

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

konten	Topik	Probabilitas
digital	9	0.137507
dukcapil	1	0.177846

Berdasarkan hasil analisis LSI *(Latent Semantic Indexing)* pada Tabel 6, terlihat bahwa beberapa kata memiliki nilai probabilitas yang mengindikasikan seberapa kuat keterkaitan kata tersebut dengan topik tertentu. Berbeda dengan LDA yang memberikan probabilitas dalam rentang positif, LSI dapat menghasilkan nilai positif dan negatif, yang menunjukkan hubungan kata dengan topik dalam ruang vektor.

Kata "dukcapil" memiliki nilai probabilitas tertinggi negatif pada topik 2 dengan -0,330946, serta pada topik 5 dengan -0,207045, mengindikasikan bahwa kata ini memiliki hubungan yang kuat dengan topik tersebut namun dalam konteks yang mungkin lebih kompleks atau beragam, seperti keluhan atau masalah terkait dengan layanan Dukcapil dalam aplikasi IKD. Di sisi lain, kata "mudah" memiliki probabilitas positif tertinggi 0,405948 pada topik 9, yang menunjukkan bahwa topik ini kemungkinan besar berkaitan dengan kemudahan penggunaan atau akses aplikasi IKD. Kata "buka" dengan probabilitas 0,190382 pada topik 0, menunjukkan bahwa kata ini cukup relevan dengan topik tersebut, kemungkinan besar terkait dengan pengalaman pengguna saat mencoba membuka aplikasi atau fitur tertentu dalam aplikasi IKD. Sementara itu, kata "wifi" dan "masuk" memiliki probabilitas positif dalam topik 8 (masing-masing 0,117265 dan 0,104064), yang menunjukkan adanya diskusi atau masalah yang berkaitan dengan konektivitas dan proses login aplikasi IKD yang menggunakan jaringan wifi.

Hasil LSI *topic probability* ini memberikan pemahaman yang lebih mendalam tentang kata-kata kunci yang relevan dengan masing-masing topik dan bagaimana mereka saling berhubungan. Analisis ini sangat berguna untuk mengidentifikasi aspek-aspek penting yang perlu diperhatikan dalam pengembangan dan peningkatan aplikasi IKD berdasarkan ulasan pengguna, serta membantu dalam memahami konteks yang lebih luas dari pengalaman pengguna.

Gambar 5. *Wordcloud* pemodelan LSI

Berdasarkan hasil visualisasi *word cloud* pada Gambar 5 dari pemodelan LSI *(Latent Semantic Indexing)*, terlihat bahwa kata-kata yang sering muncul dalam topik ulasan pengguna aplikasi Identitas Kependudukan Digital (IKD) dapat dikelompokkan ke dalam beberapa kategori utama. Beberapa kata yang menonjol seperti "aplikasi," "data," "dukcapil," "ktp," "buka," dan "digital" memiliki ukuran yang lebih besar, menunjukkan bahwa kata-kata ini memiliki frekuensi kemunculan yang tinggi dan relevan dalam berbagai topik yang diidentifikasi oleh LSI. Misalnya, kata "aplikasi" muncul sebagai kata kunci utama pada Topik 1, yang menandakan bahwa banyak ulasan pengguna berkaitan dengan pengalaman mereka dalam menggunakan aplikasi IKD.

Kata "dukcapil" dan "ktp" juga sering muncul, terutama pada topik yang terkait dengan proses verifikasi data dan layanan kependudukan, yang menunjukkan bahwa pengguna sering kali membahas masalah atau pengalaman mereka terkait integrasi dengan layanan Dinas Kependudukan dan Catatan Sipil (Dukcapil). Selain itu, kata-kata seperti "gak," "ga," "yg," dan "buka" muncul dalam beberapa topik, menandakan adanya keluhan atau masalah yang dihadapi oleh pengguna, seperti kendala saat membuka aplikasi atau masalah teknis lainnya.

Word cloud memberikan gambaran yang lebih jelas tentang topik-topik utama yang menjadi perhatian pengguna dalam aplikasi IKD. Hasil ini menunjukkan bahwa metode LSI mampu mengidentifikasi kata-kata penting yang sering muncul dalam ulasan pengguna, meskipun hubungan semantik antar kata-kata tersebut mungkin tidak sekuat LDA. Visualisasi ini sangat berguna bagi pengembang aplikasi untuk memahami aspek-aspek apa saja yang sering dibahas oleh pengguna, sehingga dapat digunakan sebagai acuan untuk meningkatkan kualitas dan kinerja aplikasi IKD.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

3.5 Evaluasi

Tabel 4. *Coherence Score*

LDA	LSI
0.4163	0.3512

Berdasarkan hasil evaluasi *Coherence Score*, metode *Latent Dirichlet Allocation* (LDA) memiliki nilai koherensi sebesar 0.4163, sedangkan *Latent Semantic Indexing* (LSI) memiliki nilai koherensi yang lebih rendah, yaitu 0.3512. *Coherence Score* merupakan metrik yang digunakan untuk mengukur seberapa koheren atau konsisten kata-kata dalam topik yang dihasilkan oleh masing-masing model. Semakin tinggi nilai *Coherence Score*, semakin baik kualitas topik yang dihasilkan karena kata-kata dalam topik tersebut memiliki keterkaitan yang lebih kuat dan relevan satu sama lain. Dalam konteks ini, LDA menunjukkan performa yang lebih baik dalam mengidentifikasi dan mengelompokkan topik utama dari ulasan pengguna aplikasi IKD dibandingkan dengan LSI.

Hasil ini mengindikasikan bahwa LDA lebih efektif dalam menangkap struktur laten dan tema yang tersembunyi dalam data teks ulasan pengguna, sehingga menghasilkan topik-topik yang lebih jelas dan terdefinisi. Sementara itu, meskipun LSI mampu menemukan hubungan semantik antar kata, hasilnya kurang koheren dibandingkan LDA. Oleh karena itu, dalam analisis *Topic Modelling* ulasan pengguna aplikasi IKD, LDA terbukti menjadi metode yang lebih unggul dan dapat memberikan gambaran yang lebih akurat tentang isu-isu dan pengalaman pengguna. Analisis ini menunjukkan bahwa pemilihan metode LDA akan memberikan hasil yang lebih relevan dan berguna bagi pengembang aplikasi untuk memahami kebutuhan dan masalah yang dihadapi oleh pengguna.

4. KESIMPULAN

Penelitian ini menganalisis perbandingan antara metode *Latent Semantic Indexing* (LSI) dan *Latent Dirichlet Allocation* (LDA) dalam *Topic Modelling* ulasan pengguna Aplikasi Identitas Kependudukan Digital (IKD) yang diperoleh dari Google Play Store. Melalui tahapan scraping data, prapemrosesan teks (termasuk *Remove Punctuation*, *Lowercase,* dan *Tokenization*), vektorisasi dengan *Bag of Words* dan DOC2BOW, hingga *Topic Modelling* menggunakan LDA dan LSI, penelitian ini berhasil mengidentifikasi tema-tema utama dalam ulasan pengguna. Hasil menunjukkan bahwa LDA menghasilkan topik-topik yang lebih koheren dan relevan dibandingkan LSI, dengan nilai *Coherence Score* 0.4163 untuk LDA dan 0.3512 untuk LSI, yang mengindikasikan efektivitas LDA dalam menangkap struktur laten data teks sehingga menghasilkan topik yang lebih jelas. Topik-topik LDA mencerminkan isu-isu nyata pengguna, seperti masalah konektivitas, fitur aplikasi, integrasi data dengan Dukcapil, dan pengalaman penggunaan secara keseluruhan. Sebaliknya, meskipun LSI mampu menemukan hubungan semantik antar kata, hasilnya kurang koheren dan tidak seefektif LDA dalam mengidentifikasi topik dari ulasan pengguna. Kesimpulannya, LDA lebih unggul dalam *Topic Modelling* untuk data ulasan aplikasi IKD dan merupakan alat yang lebih andal bagi pengembang aplikasi dalam memahami serta mengatasi permasalahan pengguna.

REFERENCES

[1] V. Salsa Bella dan D. Widodo, “Implementasi Aplikasi Identitas Kependudukan Digital (IKD) Dalam Menunjang Pelayanan Publik Masyarakat Di Kecamatan Tambaksari,” *Saraq Opat: Jurnal Administrasi Publik*, vol. 6, no. 1, hlm. 14–31, Okt 2023, doi: 10.55542/saraqopat.v6i1.833.

[2] Kementerian Dalam Negeri Republik Indonesia, “Dukcapil Terus Dukung Pengembangan IKD Menjadi INA-Pass,” https://www.kemendagri.go.id/beritaArtikel/beritakemendagri?id=36562.

[3] Muhammad Khumaidi Nursyarif, Muhamad Wahyu Tirta, Tri Wahyudi, Siti Patimah, Siti Muawwanah, dan Arbansyah Arbansyah, “Sosialisasi Identitas Kependudukan Digital Dalam Meningkatkan Partisipasi Masyarakat Kota Samarinda Pada Revolusi Digital,” *Pandawa : Pusat Publikasi Hasil Pengabdian Masyarakat*, vol. 2, no. 1, hlm. 56–63, Des 2023, doi: 10.61132/pandawa.v2i1.426.

[4] B. Setiawan, K. Ahmad Baihaqi, E. Nurlaelasari, dan H. Hikmayanti Handayani, “Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital Menggunakan Algoritma Logistic Regression dan K-Nearest Neighbor,” *Technology and Science (BITS)*, vol. 6, no. 1, hlm. 533–540, 2024, doi: 10.47065/bits.v6i1.5389.

[5] A. R. Lubis, S. Prayudani, Y. Fatmi, dan O. Nugroho, “Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) Models on News Data,” dalam *2022 5th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, Sep 2022, hlm. 314–319. doi: 10.1109/IC2IE56416.2022.9970067.

[6] D. Yamunathangam, C. B. Priya, G. Shobana, dan L. Latha, “An Overview of Topic Representation and *Topic Modelling* Methods for Short Texts and Long Corpus,” dalam *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, IEEE, Okt 2021, hlm. 1–6. doi: 10.1109/ICAECA52838.2021.9675579.

[7] U. Nur Khadijah dan N. Cahyono, “ANALISIS *TOPIC MODELLING* PARIWISATA YOGYAKARTA MENGGUNAKAN LATENT DIRICHLET ALLOCATION (LDA),” *Indonesian Journal of Computer Science*, 2024.

[8] D. Zakeshia Tiara Kannitha dan P. Kartikasari, “PEMODELAN TOPIK PADA KELUHAN PELANGGAN MENGGUNAKAN ALGORITMA LATENT DIRICHLET ALLOCATION DALAM MEDIA SOSIAL TWITTER,” vol. 11, no. 2, hlm. 266–277, 2022, [Daring]. Tersedia pada: https://ejournal3.undip.ac.id/index.php/gaussian/

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1638−1647
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5970

[9] A. Reni Dwi Astuti dan N. Cahyono, "Analisis *Topic Modelling* Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation" *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 1, hlm. 2023–326, 2023, https://doi.org/10.33022/ijcs.v12i1.3155.

[10] S. Rosales, R. Reátegui, dan C. C. Toledo, "A Topic Modeling Approach to Analyze Teaching Innovation Projects," dalam *Proceedings - 2023 4th International Conference on Information Systems and Software Technologies, ICI2ST 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, hlm. 46–53. doi: 10.1109/ICI2ST62251.2023.00014.

[11] Y. A. Singgalen, "Analisis Sentimen dan Pemodelan Topik dalam Optimalisasi Pemasaran Destinasi Pariwisata Prioritas di Indonesia," *Journal of Information Systems and Informatics*, vol. 4, no. 1, 2021.

[12] N. Cahyono, "Ekstraksi Informasi Terstruktur Profil Pengguna Website Iklan Baris" *Jurnal Buana Informatika*, vol. 12, no. 1, hlm. 39–48, 2021, doi :https://doi.org/10.24002/jbi.v12i1.4400.

[13] A. Oktavia Praneswara dan N. Cahyono, "Analisis Sentimen Ulasan Aplikasi TikTok Shop Seller Center di Google Playstore Menggunakan Algoritma Naive Bayes," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 6, hlm. 3925, 2023, doi : https://doi.org/10.33022/ijcs.v12i6.3473.

[14] Y. HaCohen-Kerner, D. Miller, dan Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, Mei 2020, doi: 10.1371/journal.pone.0232525.

[15] G. Ma, X. Wu, Z. Lin, dan S. Hu, "Drop your Decoder: Pre-training with Bag-of-Word Prediction for Dense Passage Retrieval.," dalam *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul 2024, hlm. 1818–1827. doi: 10.1145/3626772.3657792.

[16] Q. Yang, "LDA-based Topic Mining Research on China's Government Data Governance Policy," *Social Security and Administration Management*, vol. 3, no. 2, hlm. 33–42, 2022, doi: 10.23977/socsam.2022.030205.

[17] H. Yang, J. Li, dan S. Chen, "TopicRefiner: Coherence-Guided Steerable LDA for Visual Topic Enhancement," *IEEE Trans Vis Comput Graph*, vol. 30, no. 8, hlm. 4542–4557, Agu 2024, doi: 10.1109/TVCG.2023.3266890.

[18] A. S. K. Sumpter dan E. Pines, "Evaluation of Topic Models and Information Retrieval Methods in Support of Lessons Learned and Knowledge Management," dalam *2024 International Conference on System Science and Engineering (ICSSE)*, IEEE, Jun 2024, hlm. 1–6. doi: 10.1109/ICSSE61472.2024.10608962.

[19] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, dan X. Cheng, "Semantic Models for the First-Stage Retrieval: A Comprehensive Review," *ACM Trans Inf Syst*, vol. 40, no. 4, hlm. 1–42, Okt 2022, doi: 10.1145/3486250.

[20] F. Alzami *dkk.*, "[20] LDA Topic Analysis for Product Reviews in Social Media Platform," *Moneter: Jurnal Keuangan dan Perbankan*, vol. 11, no. 2, hlm. 277–283, 2023.