Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 4, August 2024, pp. 1093-1101

DOI: https://doi.org/10.52436/1.jutif.2024.5.4.1929
p-ISSN: 2723-3863
e-ISSN: 2723-3871

TEXT MINING WITH LATENT DIRICHLET ALLOCATION FOR ANALYZING PUBLIC COMMENTS ON THE M-PASSPORT APPLICATION

Theresia Shinta Hapsari*[1], Yessica Nataliani[2]

[1,2]Department of Information Systems, Faculty of Information Technology, Universitas Kristen Satya Wacana, Indonesia
Email: [1]theresiaash92@gmail.com, [2]yessica.nataliani@uksw.edu

(Article received: March 14, 2024; Revision: March 28, 2024; published: July 29, 2024)

Abstract

The M-Passport application is a service application developed by the Directorate General of Immigration of Indonesia to assist the public in applying for new passports and replacing passports online. However, in its implementation, this application has not been able to give satisfaction to its users. It is proven by the low rating of the application and the numerous negative comments on the Google Play Store. One way to identify the application's shortcomings is by analyzing user comments. In analyzing the abundance of comment data, this study utilizes the text mining method with Latent Dirichlet Allocation (LDA) topic modeling. The analysis with this method aims to find topics frequently discussed in comments so that the government can identify the shortcomings of the M-Passport application. The results of comment analysis with LDA topic modeling produced seven topics, from which three topics with the highest coherence values were selected. These three topics are then interpreted to obtain information about the public's concerns regarding the M-Passport application. The results of this interpretation include users frequently failing to log in or register to the M-Passport application, users feeling that the M-Passport application does not assist them in passport management due to constraints in the online queue feature, and some users still finding it difficult to use the M-Passport application.

Keywords: *Comments, Latent Dirichlet Allocation, Text mining, Topic modeling.*

TEXT MINING DENGAN *LATENT DIRICHLET ALLOCATION* UNTUK ANALISIS KOMENTAR MASYARAKAT TERHADAP APLIKASI M-PASPOR

Abstrak

Aplikasi M-Paspor adalah aplikasi layanan yang dikembangkan oleh Direktorat Jenderal Imigrasi Indonesia untuk membantu masyarakat dalam pangajuan permohonan paspor baru dan penggantian paspor secara online. Namun, dalam pelaksanaannya aplikasi ini belum dapat memberikan kepuasan bagi penggunanya. Hal ini dibuktikan dengan rating rendah aplikasi dan banyaknya komentar negatif di Google Play Store. Salah satu cara mengetahui kekurangan aplikasi adalah dengan melakukan analisis komentar penggunanya. Dalam menganalisis banyaknya data komentar, penelitian ini memanfaatkan metode *text mining* dengan pemodelan topik *Latent Dirichlet Allocation* (LDA). Analisis dengan metode tersebut bertujuan untuk menemukan topik yang sering menjadi pembahasan dalam komentar sehingga dapat diketahui apa saja yang menjadi kekurangan dari aplikasi M-Paspor. Hasil analisis komentar dengan pemodelan topik LDA menghasilkan tujuh topik, yang mana dari ketujuh topik tersebut dipilih tiga topik yang memiliki nilai koherensi tertinggi. Ketiga topik tersebut kemudian diinterpretasikan sehingga didapatkan informasi mengenai keresahan masyarakat terhadap aplikasi M-Paspor. Hasil interpretasi tersebut antara lain pengguna sering gagal saat masuk atau mendaftar ke aplikasi M-Paspor, pengguna merasa aplikasi M-Paspor tidak membantu mereka dalam hal pengurusan paspor karena adanya kendala pada fitur antrian online, dan sebagian pengguna masih merasa kesulitan dalam menggunakan aplikasi M-Paspor.

Kata kunci: *Komentar, Latent Dirichlet Allocation, Pemodelan topic, Text mining.*

1. PENDAHULUAN

Berkembangnya globalisasi membuat masyarakat mulai mengenal dan memanfaatkan internet untuk membantu tugas maupun aktivitas sehari-hari. Berdasarkan hasil data survei Susnas 2022, pada tahun 2022 sebanyak 66.48 % masyarakat Indonesia sudah mengakses internet [1]. Banyaknya pengguna internet ditambah dengan kemajuan bidang informasi dan teknologi menjadi alasan munculnya digitalisasi layanan guna meningkatkan efisiensi dan

efektivitas dari proses layanan. Tren digitalisasi layanan juga dimanfaatkan pemerintahan untuk melakukan fungsinya sebagai pelayan publik.

Salah satu instansi pemerintahan yang menerapkan digitalisasi layanan adalah Direktorat Jenderal Imigrasi Indonesia (Ditjen Imigrasi). Ditjen Imigrasi merupakan unsur pelaksana bidang keimigrasian yang berada di bawah dan bertanggung jawab kepada Menteri Hukum dan Hak Asasi Manusia. Salah satu layanan yang didigitalisasi adalah pengajuan paspor melalui aplikasi *mobile* yang bernama M-Paspor. M-Paspor ini bertujuan supaya memudahkan masyarakat dalam melakukan pengajuan permohonan paspor baru dan penggantian paspor secara daring melalui *smartphone*. M-Paspor sendiri sudah tersedia di App Store dan Google Play Store.

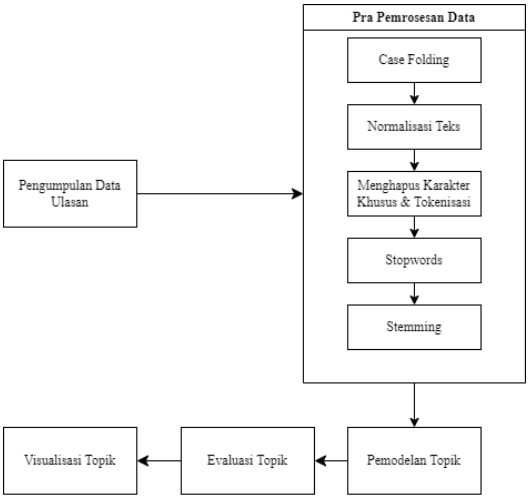
Aplikasi yang diluncurkan pada Januari 2022 ini telah diunduh lebih dari 1 juta kali di Google Play Store dan mendapatkan *rating* 1.5 dari 5 bintang berdasarkan *review* aplikasi di Google Play Store per tanggal 5 Juni 2023 [2]. Hal ini menandakan bahwa aplikasi tersebut belum mampu memberikan kepuasan bagi pengguna. Cara mengetahui alasan ketidakpuasan pengguna terhadap aplikasi adalah dengan melihat komentar yang terdapat pada bagian *Rating and reviews* di Google Play Store. Namun, banyaknya data komentar yang berjumlah ribuan memerlukan waktu lama untuk menganalisa komentar pengguna terhadap aplikasi. Oleh karena itu perlu dilakukan analisis yang lebih efisien dengan cara *text mining*. *Text mining* merupakan proses mengambil informasi yang tidak terstruktur dari teks untuk diaplikasikan dalam pekerjaan terkait *data mining* dan analisis data [3]. Terdapat dua jenis algoritma *text mining*, yaitu *supervised learning* dan *unsupervised learning* [4]. *Supervised learning* digunakan untuk mengidentifikasi kelas yang tidak dikenal dengan memberi label pada kumpulan data [5], sedangkan *unsupervised learning* merupakan suatu algoritma yang menggunakan kumpulan data input untuk mengidentifikasi pola dalam data tersebut dan menghasilkan kelompok data yang memiliki kesamaan [6].

Metode *text mining* yang dapat dimanfaatkan untuk menganalisis komentar adalah dengan *unsupervised learning* menggunakan *Latent Dirichlet Allocation* (LDA) [7]. LDA merupakan metode untuk memodelkan topik, yang mana tidak membutuhkan *training* set maupun tag khusus sehingga cocok digunakan untuk menganalisis sejumlah besar dokumen teks dalam waktu yang singkat [8]. Dibandingkan dengan algoritma ekstraksi *keyword* tradisional, LDA dapat mendeskripsikan isi dokumen lebih komprehensif dan akurat [9]. Proses generatif dalam LDA memiliki keunggulan untuk menemukan topik-topik yang tersembunyi dan memberikan wawasan baru [10].

LDA sendiri sudah banyak diaplikasikan untuk menentukan model topik pada banyak data teks maupun dokumen [11]. Sebagai contoh, penelitian terhadap persepsi pengguna internet menggunakan metode LDA dimanfaatkan untuk mengidentifikasi topik yang ramai dibahas pada laman berita detik.com. Penelitian tersebut berhasil menemukan tiga topik yang ramai dibahas dalam portal detik.com [12]. Kedua, riset yang telah dilakukan untuk mengamati tren ulasan hotel dengan pemodelan topik LDA. Riset tersebut berupaya untuk memperoleh ringkasan komentar yang dapat menjadi wawasan untuk menjaga dan mengembangkan bisnis perhotelan. Hasil yang diperoleh adalah beberapa kata yang paling sering muncul dalam topik-topik dominan, seperti lokasi, layanan, hotel, sarapan, resor, dan pantai [13]. Ketiga, penelitian terkait analisis topik yang *trending* dalam percakapan di media sosial dengan memanfaatkan LDA. Penelitian tersebut mempunyai tujuan untuk mengetahui informasi apa yang sedang dibicarakan dalam suatu grup obrolan. Penelitian tersebut menyimpulkan bahwa LDA dapat digunakan untuk mengidentifikasi topik obrolan dengan optimal, yang mana lebih dari 80% topik yang diperoleh dari sistem sesuai dengan kata-kata yang membentuk topik [14]. Terakhir, LDA juga dimanfaatkan untuk meninjau sejumlah literatur tentang virus corona. Hasil riset tersebut bertujuan untuk membantu komunitas kesehatan dan medis dalam mengekstrak informasi yang berguna dari sejumlah literatur tentang virus corona. Dengan demikian, studi tersebut dapat memberikan wawasan mengenai memahami dan mengatasi masalah yang berkaitan dengan virus corona dalam konteks kesehatan dan medis [15].

Tujuan penelitian ini adalah menganalisis data dan menyajikan hasil analisis pemodelan topik berdasarkan komentar pengguna M-Paspor. Penelitian ini diharapkan dapat menjadi informasi bagi Ditjen Imigrasi maupun pengembang aplikasi untuk memperbaiki dan mengembangkan fungsi aplikasi M-Paspor agar dapat memberikan kepuasan layanan bagi pengguna.

2. METODE PENELITIAN



Gambar 1. Diagram Tahapan Penelitian

Penelitian ini memiliki lima tahapan, yaitu pengumpulan data ulasan, pra pemrosesan data, pemodelan topik, evaluasi topik, dan visualisasi topik. Adapun *tool* yang digunakan untuk mengumpulkan data hingga visualisasi topik adalah Python. Berikut adalah tahapan penelitian yang dijelaskan pada Gambar 1.

2.1. Pengumpulan Data

Pengumpulan data dalam penelitian ini menggunakan data primer, yaitu data komentar pengguna M-Paspor pada halaman *review* di Google Play Store. Teknik yang digunakan untuk mengumpulkan data komentar tersebut adalah dengan *web scraping*. *Web scraping* merupakan cara untuk mengekstraksi data *website* dengan memanfaatkan *software* komputer. Selanjutnya data *website* tersebut akan diubah menjadi data terstruktur lalu disimpan menjadi sebuah *dataset* [16]. Penelitian ini menggunakan pemrograman Python dengan *library* BeautifulSoup untuk mengekstrak data dari *web* kemudian mengkonversi data tersebut menjadi dalam bentuk dokumen csv.

2.2. Pra Pemrosesan Data

Pra pemrosesan merupakan tahap awal dalam *text mining* [17]. Pra pemrosesan data meliputi langkah-langkah seperti menghilangkan *stopwords* (kata-kata umum dalam suatu Bahasa), menghilangkan *punctuation* atau tanda baca, memberikan label pada data, dan mentransformasikan singkatan menjadi kata sehingga dapat dengan mudah diinterpretasikan [18]. Pra pemrosesan data sendiri bertujuan untuk meningkatkan kualitas data yang akan diolah nanti.

Berikut adalah tahapan pra pemrosesan data dalam penelitian ini.

a. *Casefolding*

Casefolding merupakan proses untuk mengubah seluruh karakter huruf dalam dokumen menjadi huruf kecil (*lowercase*), seperti ”Aplikasi” menjadi ”aplikasi”, lalu ”Sulit” menjadi ”sulit”. [19].

b. Normalisasi teks

Normalisasi teks merupakan tahapan untuk mengubah kata yang tidak baku menjadi kata yang baku serta memperbaiki kesalahan ejaan penulisan dengan memanfaatkan dataset kamus normalisasi berbahasa Indonesia [20]. Berikut adalah contoh potongan dataset kamus normalisasi teks yang terdapat pada Tabel 1.

Tabel 1. Dataset Kamus Normalisasi *Slang*-Formal

Slang Words	Formal Words	Context
slalu	selalu	kode OTP sudah benar, tp slalu tidak valid.
jgn	jangan	jgn asal-asal buat aplikasi
lemot	lambat	aplikasinya lemot

c. Menghapus karakter khusus dan tokenisasi

Tahapan ini bertujuan menghapus karakter khusus karakter non ASCII, *double space*, *new line*, tanda baca (misalnya !%$#&∗?,/.;”\) dan angka, karena dinilai tidak memberikan kontribusi pada analisis data [21]. Langkah selanjutnya adalah melakukan tokenisasi, yaitu pemisahan kalimat menjadi beberapa bagian atau kata yang dapat dipahami secara gramatikal [22].

d. *Stopwords*

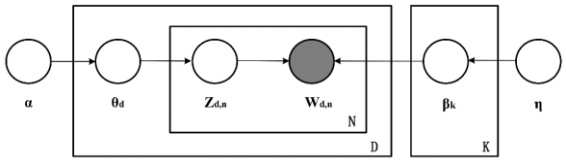
Stopwords bertujuan menghapus kata-kata umum meliputi kata sambung, kata ganti, dan kata depan karena tidak mempunyai makna untuk proses analisis teks. Tahapan ini memanfaatkan *library* NLTK dengan versi Bahasa Indonesia [23].

e. *Stemming*

Stemming merupakan proses untuk mengembalikan kata-kata menjadi bentuk dasarnya dengan menggunakan *library* Sastrawi [24]. Tujuan proses *stemming* adalah mengurangi variasi kata dalam dokumen, sehingga membantu analisis teks menjadi lebih mudah. Sebagai contoh, kata ”menggunakan” dan ”gunakan” akan distem menjadi ”guna”.

2.3. Pemodelan Topik

LDA adalah suatu pendekatan generatif yang digunakan untuk menemukan pola semantik tersembunyi dalam kumpulan dokumen besar yang relatif tidak terstruktur. Dokumen teks mengandung pola semantik tersembunyi yang disebut “topik”, dan setiap topik ini didefinisikan oleh distribusi probabilitas atas seperangkat kata yang tetap [25]. Representasi model LDA dapat dilihat pada Gambar 2.



Gambar 2. Model Representasi *Latent Dirichlet Allocation*

Berdasarkan model yang diilustrasikan pada Gambar 2, proses dalam LDA dapat dijelaskan melalui notasi berikut. Pertama, D merupakan kumpulan dokumen, K adalah kumpulan topik, sementara N merupakan jumlah kata dalam dokumen. Topik-topiknya adalah $\beta_1{:}K$, yang mana setiap β_k adalah cara distribusi kata di K. Proporsi topik untuk dokumen ke-d adalah θ_d, dimana $\theta_{d,k}$ adalah seberapa banyak topik K muncul dalam dokumen d. $Z_{d,n}$ adalah kata ke-n dalam dokumen d. Kata-kata yang teramati dalam dokumen ke-d adalah W_d, yang mana $W_{d,n}$ adalah kata ke-d dalam d. Terakhir, notasi α dan η adalah parameter *Dirichlet* [26], [27]. Rumus (1) merupakan rumus dari LDA.

$$p(w,z,\theta,\beta|\alpha,\eta) = p(\beta|\eta)p(\theta|\alpha)p(z|\theta)p(w|\beta_k) \quad (1)$$

Pada praktiknya, model LDA tidak dapat menemukan variabel z,θ,β yang tersembunyi dan

untuk menemukannya dibutuhkan estimasi parameter. Metode yang digunakan dalam estimasi parameter adalah Metode Bayesian [28][29].

Penelitian ini memanfaatkan Gensim sebagai *library* untuk pemodelan topik karena sifatnya yang *scalable* [30]. *Library* ini digunakan untuk melakukan berbagai tugas seperti, memproses teks, menghasilkan model, dan mengelola korpora teks yang besar secara efisien.

2.4. Evaluasi Topik

Tahapan untuk mengevaluasi topik hasil pemodelan LDA. Salah satu metrik yang umum digunakan adalah mengukur nilai koherensi [30]. Pengukuran nilai koherensi dipilih karena pengukuran koherensi lebih fokus pada aspek interpretatif dari suatu topik dan berkorelasi positif dengan penilaian manusia terhadap topik dibandingkan dengan metrik *perplexity*. Penelitian lain juga menyatakan jika pengukuran nilai koherensi mampu memberikan ukuran yang lebih baik untuk menilai kinerja semantik model LDA dan pemilihan topik [25].

Nilai koherensi diperoleh dengan menghitung kesamaan antar topik. Semakin kecil kesamaannya, maka semakin baik topik yang terbentuk dimana topik tersebut memiliki distribusi kata yang berbeda dengan topik-topik lainnya. Apabila semakin kecil kesamaannya, maka hasil nilai koherensinya akan semakin besar [31].

2.5. Visualisasi Topik

Tahapan untuk menyajikan hasil analisis pemodelan topik dengan menggunakan visualisasi berupa *wordcloud*. *Wordcloud* akan menampilkan susunan kata kunci dengan ukuran yang berbeda-beda yang mana semakin besar ukuran kata tersebut, maka semakin sering munculnya kata tersebut dalam suatu topik.

3. HASIL DAN PEMBAHASAN

3.1. Jumlah Data Komentar

Hasil proses dari *web scraping* terhadap ulasan pengguna M-Paspor di Google Play Store berjumlah 10.475 data. Periode data yang diambil adalah data ulasan pada 30 Desember 2021 – 28 Juni 2023. Namun, untuk mengetahui isi ulasan pengguna terhadap fitur atau perbaikan aplikasi yang terbaru, maka data akan dipilah berdasarkan tanggal terakhir aplikasi diperbarui yaitu pada 5 Maret 2023. Hasil pemilahan data tersebut memiliki total sebanyak 2.885 data.

3.2. Hasil Pra Pemrosesan Data

Data mentah kemudian diolah melalui tahap pra pemrosesan, contoh data hasil pra pemrosesan dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pra Pemrosesan Data

Process	***Result***
Raw Data	Aplikasi banyak lemot padahal pakai wifi. Isi data paling terakhir harus isi nama pasangan dan alamat pasangan. Padahal kita tidak ada pasangan. Kalau tidak isi tidak bisa click lanjut kelangka selanjutnya. Tolong perbaiki dan kami isi sembarang data jangan salahkan kami. Sekarang mau isi tanggal kedatangan sama sekali tidak bisa. Sampah tiba2 keluar sendiri.
Casefolding	aplikasi banyak lemot padahal pakai wifi. isi data paling terakhir harus isi nama pasangan dan alamat pasangan. padahal kita tidak ada pasangan. kalau tidak isi tidak bisa click lanjut kelangka selanjutnya. tolong perbaiki dan kami isi sembarang data jangan salahkan kami. sekarang mau isi tanggal kedatangan sama sekali tidak bisa. sampah tiba2 keluar sendiri.
Normalisasi Teks	aplikasi banyak lambat padahal pakai wifi. isi data paling terakhir harus isi nama pasangan dan alamat pasangan. padahal kita tidak ada pasangan. kalau tidak isi tidak bisa click lanjut kelangka selanjutnya. tolong perbaiki dan kami isi sembarang data jangan salahkan kami. sekarang mau isi tanggal kedatangan sama sekali tidak bisa. sampah tiba-tiba keluar sendiri. kalau sampai salah gunakan data2
Hapus Karakter Khusus dan Tokenisasi	[aplikasi, banyak, lambat, padahal, pakai, wifi, isi, data, paling, terakhir, harus, isi, nama, pasangan, dan, alamat, pasangan, padahal, kita, tidak, ada, pasangan, kalau, tidak, isi, tidak, bisa, click, lanjut, kelangka, selanjutnya, tolong, perbaiki, dan, kami, isi, sembarang, data, jangan, salahkan, kami, sekarang, mau, isi, tanggal, kedatangan, sama, sekali, tidak, bisa, sampah, tibatiba, keluar, sendiri]
Stopwords	[aplikasi, lambat, pakai, wifi, isi, data, isi, nama, pasangan, alamat, pasangan, pasangan, isi, click, kelangka, tolong, perbaiki, isi, sembarang, data, salahkan, isi, tanggal, kedatangan, sampah, tibatiba, salah, data]
Stemming	[aplikasi, lambat, pakai, wifi, isi, data, isi, nama, pasang, alamat, pasang, pasang, isi, click, langka, tolong, baik, isi, sembarang, data, salah, isi, tanggal, datang, sampah, tibatiba, salah, data]

Hasil pra pemrosesan data tersebut adalah data yang memiliki keseragaman karakter (*lowercase*), penulisan yang sesuai dengan kaidah Bahasa Indonesia, tidak lagi memiliki karakter-karakter khusus, lalu kalimat yang sudah dipecah menjadi per kata, tidak mengandung kata bersifat umum, dan sudah berubah menjadi kata dasar. Data ini yang selanjutnya akan menjadi kumpulan dokumen (D) dalam pemodelan topik LDA.

3.3. Hasil Pemodelan Topik

Pemodelan topik LDA dalam penelitian ini menggunakan nilai parameter jumlah topik (K) = 2-11, lalu *hyperparameter* dengan nilai $\alpha = 0.1$, dan nilai $\eta = 0.5$. Paramater α berpengaruh untuk menentukan seberapa banyak topik yang akan muncul dalam dokumen. Semakin tinggi nilai α akan menghasilkan distribusi topik yang lebih merata. Parameter η sendiri mempengaruhi bagaimana model LDA mendistribusikan kata dalam tiap topiknya. Parameter lainnya yang digunakan dalam model LDA, antara lain chunksize = 100, random_state =

100, minimum_probability = 0.05, evaluation_every = 5. Penentuan nilai *K* untuk model LDA dilakukan dengan memanfaatkan uji skor koherensi. Gambar 3 adalah potongan *script* uji pemodelan LDA dengan *library* gensim.

```
import gensim
import matplotlib.pyplot as plt
from gensim.models import LdaModel
from gensim.models.coherencemodel import CoherenceModel

# Data corpus dan dictionary
corpus = doc_term_matrix
dictionary = dictionary

# Range nilai num_topics yang akan diuji
num_topics_range = range(2, 11)

# Daftar untuk menyimpan skor koherensi
coherence_scores = []

# Melakukan iterasi untuk berbagai nilai num_topics
for num_topics in num_topics_range:
    # Membangun model LDA
    lda_model = LdaModel(
        corpus=corpus,
        id2word=dictionary,
        num_topics=num_topics,
        random_state=100,
        chunksize=100,
        passes=10,
        alpha=0.1,
        eta=0.5,
        eval_every=5,
        minimum_probability=0.05,
        per_word_topics=True
    )

    # Menghitung skor koherensi
    coherence_model = CoherenceModel(model=lda_model,
        texts=df_cleaned['textdata_tokens_stemmed2'],
        dictionary=dictionary, coherence='c_v')
    coherence_score = coherence_model.get_coherence()
    coherence_scores.append(coherence_score)

# Membuat plot untuk skor koherensi
plt.figure(figsize=(10, 6))
plt.plot(num_topics_range, coherence_scores, marker='o',
         linestyle='-', color='b')
plt.title('Grafik Skor Koherensi untuk LDA')
plt.xlabel('Jumlah Topik')
plt.ylabel('Skor Koherensi')
plt.xticks(num_topics_range)
plt.grid(True)
plt.show()
```

Gambar 3. *Script* Python Pemodelan Topik LDA Menggunakan *Library* Gensim

Adapun hasil pengujian yang terdapat di Gambar 3 tersebut dapat dilihat pada grafik yang ditampilkan di Gambar 4. Berdasarkan Gambar 4, jumlah topik yang memiliki nilai koherensi tertinggi terletak pada titik dengan jumlah topik (*K*) sebanyak tujuh, dengan nilai koherensi sebesar 0.475. Nilai koherensi yang tinggi menandakan bahwa model LDA tersebut sudah optimal, sehingga penelitian ini menetapkan *K* = 7 sebagai jumlah topik hasil pemodelan LDA.

Gambar 4. Grafik Skor Koherensi untuk Jumlah Topik

Proses pemodelan topik menghasilkan distribusi topik per dokumen. Berikut adalah potongan hasil distribusi topik dari beberapa dokumen yang ditampilkan dalam Tabel 3.

Tabel 3. Hasil Distribusi Topik per Dokumen untuk Tiga Dokumen Pertama

Dokumen	Topik	Probabilitas
1	6	0.872
	1	0.182
2	4	0.568
	5	0.190
3	1	0.408
	2	0.407

Tabel 3 menampilkan beberapa nilai probabilitas suatu dokumen berkaitan dengan topik tertentu. Semakin tinggi nilai probabilitasnya, semakin kuat juga hubungan dokumen dengan topik tersebut. Sebagai contoh untuk Dokumen 2 yang memiliki nilai probabilitas paling tinggi dengan Topik 4, dibandingkan dengan Topik 1 dan Topik 5. Nilai probabilitas yang paling tinggi menandakan bahwa Topik 4 lebih mendominasi dalam Dokumen 2 dibandingkan dengan Topik 1 maupun Topik 5. Adapun hasil distribusi kata untuk tiap topiknya disajikan pada Tabel 4.

Tabel 4. Hasil Distribusi Kata per Topik

Topik (*K* = 7)	***Words & Score of Importance***			
1	tanggal (0.054)	pilih (0.045)	…	menu (0.0063)
2	masuk (0.433)	coba (0.392)	…	no (0.0064)
3	imigrasi (0.054)	buka (0.049)	…	daftar (0.0051)
4	jelek (0.032)	negara (0.025)	…	pajak (0.0042)
5	out (0.071)	time (0.067)	…	enak (0.0039)
6	lot (0.094)	again (0.017)	…	du (0.0023)
7	aplikasi (0.187)	error (0.054)	…	indonesia (0.0071)

Tabel 4 menyajikan hasil distribusi kata per topik yang mempunyai rentang nilai 0 – 1. Nilai-nilai tersebut menunjukkan seberapa besar peluang munculnya kata tersebut dalam suatu topik [32]. Semakin tinggi probabilitasnya, maka kata tersebut dianggap semakin penting atau relevan kata-kata tersebut dengan topik.

3.4. Evaluasi Topik

Langkah selanjutnya adalah menilai apakah suatu kelompok topik sudah baik atau belum dengan melakukan perhitungan koherensi terhadap tiap topik. Berikut adalah tabel hasil perhitungan nilai koherensi untuk tiap topiknya.

Tabel 5. Nilai Koherensi untuk Tiap Topik

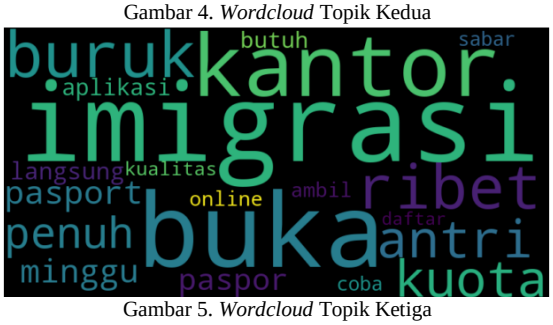
Topik	***Coherence Score***
1	0.495
2	0.694
3	0.616
4	0.401
5	0.422
6	0.484
7	0.520

Berdasarkan data pada Tabel 5 tiga nilai koherensi tertinggi berada pada kelompok topik kedua dengan nilai 0.694, kelompok topik ketiga dengan nilai 0.616, dan kelompok ketujuh dengan nilai 0.520. Nilai koherensi yang semakin tinggi dapat memudahkan proses penginterpretasian makna berdasarkan kumpulan kata kunci yang menyusunnya [27].

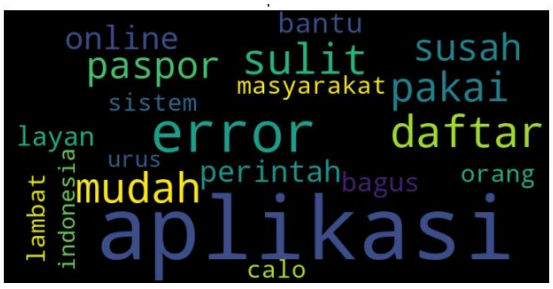
3.5. Visualisasi Topik

Tahap berikutnya hasil pemodelan topik akan ditampilkan dalam visualisasi berbentuk *wordcloud*. Visualisasi ini mencakup 20 kata dominan untuk setiap topiknya. Berikut adalah hasil visualisasi terhadap topik kedua yang dapat dilihat pada Gambar 4, kemudian topik ketiga yang ditampilkan pada Gambar 5, dan Gambar 6 yang menampilkan topik ketujuh.

Gambar 4. *Wordcloud* Topik Kedua



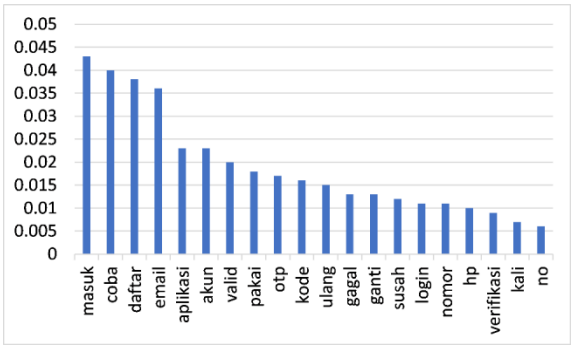
Gambar 5. *Wordcloud* Topik Ketiga



Gambar 6. *Wordcloud* Topik Ketujuh

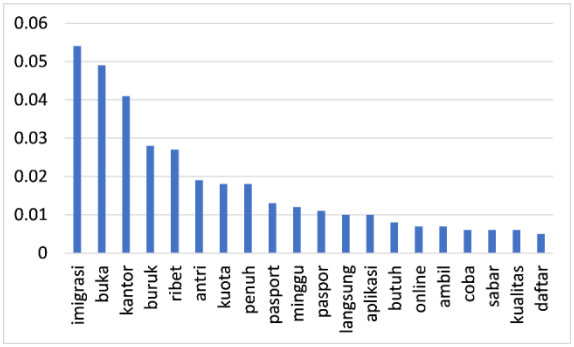
Kata-kata yang muncul pada hasil visualisasi yang terdapat pada Gambar 4, Gambar 5, dan Gambar 6 merupakan kata-kata yang memiliki probabilitas kemunculan paling tinggi dalam suatu topik. Untuk dapat memberikan informasi, sekumpulan kata dominan tersebut memerlukan proses interpretasi makna oleh manusia. Interpretasi makna tersebut dilakukan dengan membandingkan hasil pemodelan topik dengan data aslinya.

Gambar 7 sampai Gambar 9 menunjukkan hasil daftar kata dari ketiga topik yang diurutkan berdasarkan frekuensi kemunculan kata tersebut.



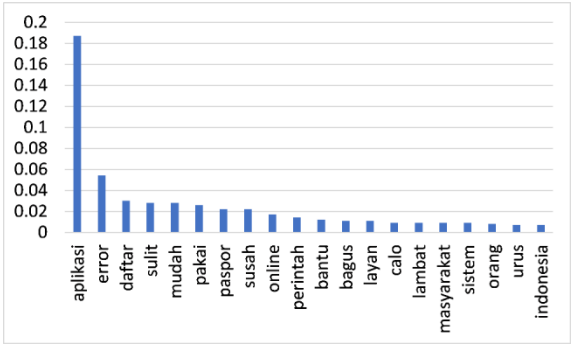
Gambar 7. Daftar Kata pada Topik Kedua

Daftar kata pada Gambar 7 meliputi ”masuk”, ”coba”, ”daftar”, ”*email*”, ”aplikasi”, ”akun”, ”susah”, ”gagal". Daftar kata tersebut dapat diinterpretasikan menjadi pengguna masih mengalami kendala saat akan masuk ataupun mendaftarkan akun ke aplikasi M-Paspor.



Gambar 8. Daftar Kata pada Topik Ketiga

Berikutnya pada Gambar 8 yang memuat daftar kata dominan untuk topik ketiga memiliki daftar kata sebagai berikut ”imigrasi”, ”buka”, ”kantor”, ”buruk”, ”ribet”, ”antri”, ”penuh”, “paspor”, “langsung”, “aplikasi”, “*online*”. Kumpulan kata kunci tersebut dapat diartikan menjadi keluhan pengguna yang merasa tidak terbantu dengan aplikasi karena mengalami kendala saat menggunakan fitur antrian *online*.



Gambar 9. Daftar Kata pada Topik Ketujuh

Terakhir topik ketujuh yang ada pada Gambar 9 memiliki kumpulan kata kunci yang terdiri dari

”aplikasi”, ”*error*”, ”daftar”, ”sulit”, ”mudah”, ”pakai”, ”paspor”, ”susah”. Kata-kata kunci tersebut dapat diinterpretasikan bahwa sebagian orang masih merasa kesulitan menggunakan aplikasi M-Paspor.

4. DISKUSI

Berdasarkan riset terdahulu yang dilakukan oleh Yayang M., Junaidi, dan Iman S. (2023) tentang pemodelan topik pada judul berita *online* detikcom menggunakan *Latent Dirichlet Allocation*”, bertujuan untuk mengetahui beberapa topik dari data berita. Jumlah topik yang dihasilkan adalah tiga dengan nilai koherensi 0.7586 [19].

Riset selanjutnya tentang analisis pemodelan topik ulasan aplikasi BNI, BCA, dan BRI menggunakan *Latent Dirichlet Allocation* yang diteliti oleh Benny A.T, dkk (2023). Tujuan dari penelitian ini adalah menggunakan metode LDA untuk mengelompokkan ulasan menjadi topik yang mampu menggambarkan pengalaman pengguna. Pemodelan topik tersebut memberikan temuan tiga topik utama untuk masing-masing aplikasi [33].

Penelitian sebelumnya hanya berfokus dengan penggunaan uji nilai koherensi untuk menentukan jumlah topik yang optimal bagi model LDA. Dalam penelitian ini uji nilai koherensi dilakukan sebanyak dua kali, yang pertama untuk mencari jumlah topik yang dapat memberikan nilai koherensi tertinggi bagi suatu model, dan kedua uji nilai koherensi per topik untuk menemukan topik yang maknanya mudah untuk diinterpretasikan manusia. Hasil dari penelitian ini adalah tujuh topik dengan nilai koherensi sebesar 0.475, yang difokuskan lagi menjadi tiga topik utama berdasarkan urutan nilai koherensi yang tertinggi.

5. KESIMPULAN

Penelitian ini telah mampu menghasilkan analisis topik ulasan pengguna terhadap aplikasi M-Paspor menggunakan pemodelan topik LDA. Terdapat tujuh topik dengan 20 kata kunci pada tiap topiknya yang dihasilkan dalam proses pemodelan topik LDA. Dari ketujuh topik tersebut diambil tiga topik utama yang memiliki nilai koherensi lebih tinggi dibandingkan dengan topik lainnya. Topik tersebut diantaranya adalah Topik 2, Topik 3, dan Topik 7. Hasil interpretasi dari ketiga topik adalah Topik 2 berisikan informasi mengenai pengguna yang sering mengalami gagal saat mencoba masuk maupun mendaftar ke aplikasi, Topik 3 menghasilkan informasi keluhan pengguna yang merasa tidak terbantu dengan aplikasi karena ada kendala saat mendaftar antrian online, dan Topik 7 berisikan informasi bahwa masih terdapat Sebagian orang yang merasa kesulitan menggunakan layanan aplikasi M-Paspor.

Hasil analisis tersebut dapat menjadi masukan bagi pihak Ditjen Imigrasi untuk dapat meningkatkan fungsi aplikasi M-Paspor agar dapat memberikan kepuasan layanan bagi masyarakat yang akan mengurus paspor. Penelitian ini bisa diperdalam lagi dengan melakukan analisis yang berfokus terhadap kritik dari aplikasi M-Paspor agar dapat mengetahui kekurangan aplikasi secara lebih rinci.

DAFTAR PUSTAKA

[1] Badan Pusat Statistik Republik Indonesia, *Statistik Telekomunikasi Indonesia 2022*. Jakarta: Badan Pusat Statistik, 2023.

[2] Google, “M-Paspor,” Google Play Store. Accessed: Jun. 28, 2023. [Online]. Available: https://play.google.com/store/apps/details?id=id.go.imigrasi.paspor_online&hl=id

[3] Y. Kalepalli, P. D. P. Teja, S. Tasneem, and S. Manne, “Effective Comparison of LDA with LSA for Topic Modelling,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)*, 2020.

[4] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, “Text Mining in Big Data Analytics,” *Big Data Cogn. Comput.*, vol. 4, no. 1, pp. 1–34, 2020, doi: 10.3390/bdcc4010001.

[5] A. Roihan, P. A. Sunarya, and A. S. Rafika, “Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper,” *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[6] M. N. Faiz, O. Somantri, A. R. Supriyono, and A. W. Muhammad, “Impact of Feature Selection Methods on Machine Learning-based for Detecting DDoS Attacks : Literature Review,” *J. Informatics Telecommun. Eng.*, vol. 5, no. 2, pp. 305–314, 2022, doi: 10.31289/jite.v5i2.6112.

[7] V. Vukanti and A. Jose, “Business Analytics: A Case-study Approach Using LDA Topic Modelling,” in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, IEEE, 2021, pp. 1818–1823. doi: 10.1109/ICCMC51019.2021.9418344.

[8] F. Gurcan, O. Ozyurt, and N. E. Cagiltay, “Investigation of Emerging Trends in the E-Learning Field Using Latent Dirichlet Allocation,” *Int. Rev. Res. Open Distance Learn.*, vol. 22, no. 2, pp. 1–18, 2021, doi: 10.19173/irrodl.v22i2.5358.

[9] Y. Zhang and L. Zhang, “Movie Recommendation Algorithm Based on Sentiment Analysis and LDA,” in *The 8th International Conference Technology and Quantitative Management (ITQM 2020 & 2021)*, Elsevier B.V., 2021, pp. 871–878. doi: 10.1016/j.procs.2022.01.109.

[10] J. H. Lee and M. J. Ostwald, "Latent Dirichlet Allocation (LDA) Topic Models for Space Syntax Studies on Spatial Experience," *City, Territ. Archit.*, vol. 11, no. 3, pp. 1–20, 2024, doi: 10.1186/s40410-023-00223-3.

[11] Y. Sahria and D. Hatta Fudholi, "Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation)," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 4, no. 2, pp. 336–344, 2020.

[12] A. R. D. Astuti and N. Cahyono, "Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 326–334, 2023, doi: 10.33022/ijcs.v12i1.3155.

[13] S. Suparyati, E. Utami, and A. Fathurahman, "Pengamatan Tren Ulasan Hotel Menggunakan Pemodelan Topik Berbasis Latent Dirichlet Allocation," *J. Appl. Informatics Comput.*, vol. 6, no. 1, pp. 71–77, 2022, doi: 10.30871/jaic.v6i1.3645.

[14] A. Syaifuddin, R. A. Harianto, and J. Santoso, "Analisis Trending Topik untuk Percakapan Media Sosial dengan Menggunakan Topic Modelling Berbasis Algoritme LDA," *INSYST J. Intell. Syst. Comput.*, vol. 2, no. 1, pp. 12–19, 2020, doi: https://doi.org/10.52985/insyst.v2i1.150.

[15] X. Cheng, Q. Cao, and S. S. Liao, "An Overview of Literature on COVID-19, MERS and SARS: Using Text Mining and Latent Dirichlet Allocation," *J. Inf. Sci.*, vol. 48, no. 3, pp. 304–320, 2022, doi: 10.1177/0165551520954674.

[16] M. A. Khder, "Web Scraping or Web cRawling: State of Art, Techniques, Approaches and Application," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[17] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.

[18] M. Nesca, A. Katz, C. K. Leung, and L. M. Lix, "A Scoping Review of Preprocessing Methods for Unstructured Text Data to Assess Data Quality," *Int. J. Popul. Data Sci.*, vol. 7, no. 1, 2022, doi: 10.23889/ijpds.v6i1.1757.

[19] Y. Matira, Junaidi, and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi J. Stat. Its Appl.*, vol. 4, no. 1, pp. 2721–379, 2023, doi: 10.20956/ejsa.vi.24843.

[20] D. Ferarizki, M. Fikry, F. Yanto, and F. Insani, "Klasifikasi Sentimen Masyarakat di Twitter Terhadap Ancaman Resesi Ekonomi 2023 dengan Metode K-Nearest Neighbor," *Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 2, pp. 1111–1120, 2023, doi: 10.30865/klik.v4i2.1315.

[21] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints," *Expert Syst. Appl.*, vol. 127, pp. 256–271, Aug. 2019, doi: 10.1016/j.eswa.2019.03.001.

[22] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, "Opportunities and Challenges of Text Mining in Materials Research," *iScience*, vol. 24, no. 3, 2021, doi: 10.1016/j.isci.2021.102155.

[23] D. L. C. Pardede and M. A. I. Waskita, "Analisis Pemodelan Topik untuk Ulasan Tentang Peduli Lindungi," *J. Ilm. Inform. Komput.*, vol. 28, no. 1, pp. 17–26, 2023, doi: 10.35760/ik.2023.v28i1.7925.

[24] R. Prabowo, H. Sujaini, and T. Rismawan, "Analisis Sentimen Pengguna Twitter Terhadap Kasus COVID-19 di Indonesia Menggunakan Metode Regresi Logistik Multinomial," *J. Sist. dan Teknol. Inf.*, vol. 11, no. 1, pp. 85–90, Jan. 2023, doi: 10.26418/justin.v11i1.57450.

[25] S. Zhou, P. Kan, Q. Huang, and J. Silbernagel, "A Guided Latent Dirichlet Allocation Approach to Investigate Real-time Latent Topics of Twitter Data During Hurricane Laura," *J. Inf. Sci.*, vol. 49, no. 2, pp. 465–479, Apr. 2023, doi: 10.1177/01655515211007724.

[26] D. Maulidiya, "Topic Modelling using Latent Dirichlet Allocation (LDA) to Investigate the Latent Topics of Mathematical Creative Thinking Research in Indonesia," *J. Intell. Comput. Heal. Inform.*, vol. 3, no. 2, pp. 34–35, 2022, doi: 10.26714/jichi.v3i2.11428.

[27] N. L. P. Merawati, A. Z. Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, Feb. 2021, doi: 10.29207/resti.v5i1.2587.

[28] U. T. Setijohatmo, S. Rachmat, T. Susilawati, and Y. Rahman, "Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik," in *Prosiding The 11th Industrial Research Workshop and National*

Seminar, Bandung, 2020, pp. 402–408.

[29] D. Z. T. Kannitha, M. Mustafid, and P. Kartikasari, "Pemodelan Topik pada Keluhan Pelanggan Menggunakan Algoritma Latent Dirichlet Allocation dalam Media Sosial Twitter," *J. Gaussian*, vol. 11, no. 2, pp. 266–277, 2022, doi: 10.14710/j.gauss.v11i2.35474.

[30] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, "Topic modeling: Comparison of LSA and LDA on Scientific Publications," in *2021 4th International Conference on Data Storage and Data Engineering (DSDE'21)*, Association for Computing Machinery, Feb. 2021, pp. 59–64. doi: 10.1145/3456146.3456156.

[31] A. Muhaimin *et al.*, "Social Media Analysis and Topic Modeling: Case Study of Stunting in Indonesia," *J. Inform. dan Teknol. Inf.*, vol. 20, no. 3, pp. 406–415, 2023, doi: 10.31515/telematika.v20i3.10797.

[32] N. A. Sanjaya, "lmplementasi Latent Dirichlet Allocation (LDA) untuk Klasterisasi Cerita Berbahasa Bali," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, pp. 127–134, 2021, doi: 10.25126/jtiik.202183556.

[33] B. A. Tondang, M. R. Fadhil, M. N. Perdana, A. Fauzi, and U. S. Janitra, "Analisis Pemodelan Topik Ulasan Aplikasi BNI, BCA, dan BRI Menggunakan Latent Dirichlet Allocation," *INFOTECH J. Inform. Teknol.*, vol. 4, no. 1, pp. 114–127, 2023, doi: 10.37373/infotech.v4i1.601.