Jurnal Ilmiah Informatika Komputer, 28(1): 17-26.
Doi: https://doi.org/10.35760/ik.2023.v28i1.7925

ANALISIS PEMODELAN TOPIK UNTUK ULASAN TENTANG PEDULI LINDUNGI

[1]D. L. Crispina Pardede, [2]Muhammad Andrias Indra Waskita*
[1,2]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
[1]pardede@staff.gunadarma.ac.id, [2]indra007@student.gunadarma.ac.id
**) Penulis korespondensi*

Abstrak

Pandemi COVID-19 yang berlangsung sejak awal tahun 2020 membuat berbagai negara menerapkan kebijakan protokol kesehatan mulai dari pencegahan, isolasi, hingga perawatan pasien. Pemerintah Republik Indonesia mengambil langkah pencegahan dengan mengembangkan sebuah aplikasi berbasis mobile dengan nama "PeduliLindungi". Keberadaan aplikasi tersebut banyak mendapat reaksi dari masyarakat. Berbagai ulasan diunggah melalui berbagai media daring. Penelitian ini melakukan analisis pemodelan topik untuk ulasan-ulasan yang diunggah melalui Google Play Store. Metode yang diterapkan adalah metode Latent Dirichlet Allocation (LDA). Pemodelan topik menghasilkan nilai coherence sebesar 0.3963 untuk jumlah topik sebanyak 5.

Kata Kunci: *Latent Dirichlet Allocation, Natural Language Processing, PeduliLindungi, Pemodelan Topik*

Abstract

The COVID-19 pandemic, which has been going on since early 2020, has forced various countries to implement health protocol policies ranging from prevention, isolation to patient care. The Government of the Republic of Indonesia took a preventive step by developing a mobile-based application with the name "PeduliLindungi". The existence of this application received a lot of reactions from the public. Various reviews uploaded through various online media. This study conducts a topic modeling analysis for reviews uploaded through the Google Play Store. The method applied is the Latent Dirichlet Allocation (LDA) method. The topic modelling resulted a coherence score of 0.3963 for a total of 5 topics.

Keywords: *Latent Dirichlet Allocation, Natural Language Processing, PeduliLindungi, Topic Modelling*

PENDAHULUAN

Pada akhir tahun 2019, ditemukan jenis penyakit baru yang disebabkan oleh virus corona yang kemudian disebut dengan COVID-19. Virus ini menyebar dengan cepat dari satu kota di negara China ke hampir seluruh dunia. Indonesia merupakan salah satu negara yang terkena infeksi penyebaran virus corona. Untuk mengendalikan penyebaran virus ini, diperlukan tindakan oleh pemerintah dan kesadaran penuh oleh masyarakat. Pemerintah, dalam hal ini Kemenkominfo meresmikan Aplikasi PeduliLindungi pada tahun 2020. Aplikasi tersebut merupakan gagasan Kominfo, Gugus Tugas Covid-19 yang mengkoordinasikan Kementerian BUMN, BNPB, Kementerian Kesehatan, TNI,

Polri, dan Kementerian Aparatur Negara dan Reformasi Birokrasi. Aplikasi Pedulilindungi dibangun Pendayagunaan dalam rangka pelaksanaan surveilans kesehatan penanganan Corona Virus Disease 2019 (COVID-19) dan ditetapkan dalam Keputusan Menteri Kominfo Nomor 171 Tahun 2020 [1].

Aplikasi PeduliLindungi ditetapkan sebagai yang dipergunakan dalam pelaksanaan surveilans kesehatan oleh Pemerintah dalam mengangani penyebaran COVID-19, antara lain penelusuran (*tracing*); pelacakan (*tracing*); dan pemberian peringatan (*warning* dan *fencing*) [2]. Penambahan fitur kemudian dilakukan berdasarkan Keputusan Menteri Kominfo Nomor 253 Tahun 2020 [3], sehingga Aplikasi PeduliLindungi memiliki fitur berupa penelusuran (*tracing*); pelacakan (*tracking*); pemberian peringatan (*warning* dan *fencing*); e-sertifikat yang meliputi hasil tes Rapid Test dan/atau Swab Test, Surat Keterangan Sehat, Surat Keterangan Sembuh Covid19, Surat keterangan vaksinasi, Surat Izin Keluar/Masuk, Surat Penugasan Instansi, dan/atau sertifikat kesehatan lainnya; Sistem Pemosisi Global (GPS); Catatan Harian Digital (*digital diary*); dan/atau fitur lain yang ditetapkan dan/atau kerja sama dengan platform lain.

Total pengguna aplikasi Peduli-Lindungi, bulan Juni 2020 tercatat mencapai 4.025.861 jiwa [4]. Jumlah unduhan aplikasi PeduliLindungi mencapai lebih dari 50.000.000 unduhan di Google Play Store. Rating aplikasi PeduliLindungi adalah 4,4 menunjukkan aplikasi termasuk dalam kategori disukai oleh pengguna. Berbagai ulasan tentang aplikasi PeduliLindungi diberikan melalui Google Play Store. Ulasan yang disampaikan meliputi beragam hal. Adalah hal yang menarik untuk mengetahui topik-topik yang disinggung di dalam berbagai ulasan. Topik-topik yang teridentifikasi dapat digunakan sebagai masukan, khususnya bagi pemerintah, untuk mengoptimalisasi aplikasi PeduliLindungi. Penelitian ini melakukan pemodelan topik untuk mengetahui topik apa saja yang ada dalam berbagai ulasan tentang PeduliLindungi yang disampaikan melalui Google Play Store.

Pemodelan topik atau *Topic Modeling* adalah satu metode *unsupervised machine learning* untuk mengorganisasi teks [5]. Algoritma pemodelan topik menyediakan teknik untuk mengelompokkan tema sebagai topik [6]. Pemodelan topik termasuk dalam *soft/fuzzy clustering* yang mana setiap objek dapat dimiliki lebih dari satu *cluster*. Terdapat beberapa teknik yang digunakan untuk pemodelan topik, salah satunya adalah metode *Latent Dirichlet Allocation* (LDA) [7][8]. LDA adalah metode pemodelan topik yang dapat mengklasifikasikan sebuah teks dalam suatu dokumen yang sangat besar ke topik tertentu LDA juga digunakan untuk meringkas, mengelompokkan, menghubung-kan dan memproses data.

Berbagai penelitian mengenai pemodelan dilakukan menggunakan LDA.

Nugraha dan Munggaran dalam penelitiannya menggunakan LDA untuk melakukan pemodelan topik terhadap teks berita [9]. LDA juga diterapkan untuk Pemodelan Topik Temu Kembali Informasi dalam Rekomendasi Tugas Akhir [10], pemodelan topik skripsi [11], Ekstraksi Berita Saham Online [12], dan Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara [13].

Penelitian ini melakukan analisis pemodelan topik terhadap ulasan para pengguna aplikasi PeduliLindungi yang diambil melalui Google Play Store. Metode LDA diterapkan untuk mengelompokkan ulasan yang diberikan oleh pengguna aplikasi PeduliLindungi. Pemodelan topik mengidentifikasi topik-topik yang pada ulasan mengenai aplikasi Peduli Lindungi

METODE PENELITIAN

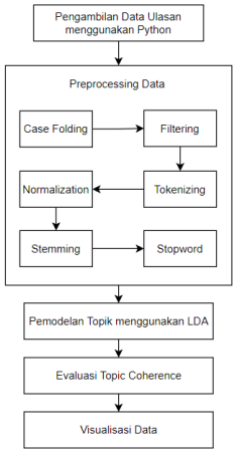
Data berupa ulasan pengguna PeduliLindungi diambil dari Google Play Store menggunakan sebuah program yang disusun dalam bahasa python. Sebanyak 20.000 ulasan berhasil dikumpulkan menggunakan program tersebut.

Guna mendapatkan data teks yang terstruktur, data yang terkumpul diolah pada tahap pra-proses teks (*text preprocessing*), yang meliputi langkah-langkah *case folding*, *filtering*, *tokenizing*, *normalization*, *stemming*, dan *stopword*. Penerapan metode LDA dilakukan kepada teks yang telah terstruktur, dilanjutkan dengan evaluasi *topic coherence* dan visualisasi data (Gambar 1).

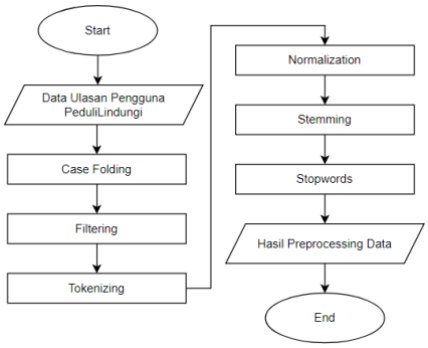
Text Preprocessing

Pada tahap *preprocessing* dilakukan proses pembersihan data yang sebelumnya tidak terstruktur menjadi data teks yang terstruktur. Tahap *preprocessing* [14][15][16] dilakukan dalam rangkaian langkah *case folding, filtering, tokenizing, normalization, stemming* dan *stopwords* (Gambar 2).

1. *Case Folding*: Pengubahan karakter huruf besar ‘A’ hingga ‘Z’ menjadi karakter ‘a’ hingga ‘z’. Karakter-karakter selain ‘a’ hingga ‘z’ seperti tanda baca dan angka dihilangkan.
2. *Filtering*: Penghilangan baris baru, angka, emoji, tautan, dan simbol seperti (,),$,*,@,?,!,(.) pada setiap data ulasan.
3. *Tokenizing*: Pemecahan kalimat menjadi kumpulan kata.
4. *Normalization*: Normalisasi kata agar data ulasan lebih terstruktur, dan telah dibuat kamus normalisasi sebanyak 277 kata.
5. *Stemming*: Penghilangan semua imbuhan yang ada pada kata sehingga menjadi kata dasar.
6. *Stopwords*: Penghilangan kata yang memiliki fungsi tetapi tidak memiliki makna/arti, seperti kata sambung, kata depan, dan kata ganti. Kamus stopwords yang digunakan merupakan kamus *stopword yang* berasal dari *library* NLTK berbahasa Indonesia.



Gambar 1. Tahap Penelitian



Gambar 2. Tahap *Preprocessing*

Metode *Latent Dirichlet Allocation*

Metode LDA berperan dalam menemukan topik yang ada di dalam sebuah dokumen dengan ukuran yang sangat besar. Topik ditemukan berdasarkan frekuensi kata yang sering muncul.

Urutan langkah dalam melakukan proses LDA: meng-*import* data set yang telah melalui *text preprocessing.* Pengklasifikasian beberapa kata yang mirip ke dalam topik dengan memanfaatkan fungsi *lda.model* dari *library gensim*.

Penentuan setiap kata yang masuk ke dalam topik tertentu, sebut topik *k*, dilakukan melalui proses distribusi probabilitas yang memanfaatkan Gibbs Sampling (1). Gibbs Sampling digunakan untuk mengestimasi probabilitas topik tertentu (topik *k*) terhadap kata dan dokumen terhadap topik *k* tersebut.

$$P(z_i = j \mid z_{-i}, w_i, d_i, .) \propto \frac{C_{wij}^{WT} + \beta}{\sum_{W=1}^{W} C_{wij}^{WT} + W\,\beta} \frac{C_{dij}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dit}^{DT} + T\alpha} \quad (1)$$

Dimana *i* merupakan kata dan *j* merupakan dokumen; z_i adalah kata yang diwakili dengan variabel *z*; z_{-i} adalah kata selain kata *z*; w_i adalah jumlah kata; d_i adalah kata yang di- *assign* ke dokumen yang diwakili dengan variabel *d*. *WT* adalah kata dalam topik ; *DT* adalah topik dalam dokumen; *wij* adalah jumlah kata di dokumen; *wj* jumlah kata

di selain dokumen; *dij* adalah kata yang di dokumen *d*; *dit* adalah topik tersembunyi dalam dokumen; *dj* adalah dokumen selain dokumen *d*. C_{wij}^{WT} menunjukkan berapa kali *w* di-*assign* ke topik *k* di setiap dokumen, C_{dij}^{DT} berapa kali topik *k* tersebut di-*assign* ke dokumen *d*, C_{wj}^{WT} berapa kali *w* di-*assign* ke topik selain topik *k* di setiap dokumen, C_{dj}^{DT} berapa kali topik *k* di-*assign* ke selain *d*, α adalah parameter *dirichlet* atas distribusi dokumen terhadap topik, diambil dari 50/*T*, nilai tersebut adalah nilai standar untuk distribusi dirichlet, β adalah parameter *dirichlet* atas distribusi topik terhadap kata di semua dokumen yaitu 0.001, nilai tersebut adalah nilai standar untuk distribusi *dirichlet*, W adalah jumlah kata untuk setiap dokumen dan T adalah jumlah topik untuk setiap dokumen.

Evaluasi Topic Coherence

Topic coherence merupakan satu set dari kata-kata yang didapat pada topik model yang dinilai sesuai dengan tingkat koherensi atau dalam diinterpretasi oleh manusia dengan tingkat kemudahannya.

Topic coherence mengukur nilai suatu topik dengan mengukur kesamaan semantik antar kata yang berada dalam topik. Pengukuran ini dilakukan untuk membantu membedakan antara topik yang didapat secara semantic dengan topik yang memiliki keterkaitan secara statistik. *Topic coherence* adalah suatu ukuran yang digunakan untuk mengevaluasi pemodelan topik. Jika *coherence score* pada suatu topik bernilai tinggi, maka model yang dihasilkan dapat dianggap baik.

Pada penelitian ini, *coherence score* digunakan untuk mengevaluasi nilai koherensi jumlah topik sebanyak 1 sampai dengan topik sebanyak 10, kemudian dibandingkan dengan nilai koherensi pada topik sebanyak 5 yang sudah dihasilkan menggunakan *coherence score*.

Rumus perhitungan *coherence score* dapat dilihat sebagai berikut:

$$score(vi, vj, \in) = log \frac{D(vi,vj)+ \in}{D(vj)} \quad (2)$$

Dimana $score(vi, vj, \in)$ adalah *coherence score* untuk kata $vi\ dan\ vj$, $D(vi, vj)$ adalah jumlah dokumen yang memuat kata vi dan vj, $D(vj)$ adalah jumlah dokumen yang memuat kata vj, vi adalah frekuensi kata vi, vj adalah frekuensi kata vj, $\in$ adalah variabel yang menjamin hasil bernilai positif.

Visualisasi Data

Visualisasi data adalah hasil dari analisis pemodelan topik ulasan pengguna PeduliLindungi menggunakan *library gensim* pada bahasa pemrograman python.

Hasil visualisasi data adalah berupa *Intertropic Distance Map.* Program mengambil data hasil proses LDA, kemudian program menggunakan pyLDAvis dan library gensim yang ada pada kode program berbahasa python untuk membentuk visualisasi. Proses akhir proses, program akan menghasilkan visualisasi berupa grafik *Intertropic Distance Map*.

HASIL DAN PEMBAHASAN

LDA memproses data teks dengan mengumpulkan dokumen dengan topik yang mirip, kemudian diklasifikasikan ke kelompok kata yang sering muncul dan menghasilkan topik melalui distribusi probabilitas atas kata-kata tersebut.

Gambar 3 menunjukkan hasil dari pembentukan topik dengan 10 kata kunci teratas untuk ditampilkan nilai bobotnya. Hasil *output* berupa 5 macam topik yang telah selesai dibentuk, dengan setiap topiknya terdapat 10 kata kunci teratas atau kata yang paling sering muncul di setiap topiknya untuk mewakili topik tersebut. Pada topik 1 terdapat 10 macam perwakilan kata kunci, setiap kata kunci memiliki nilai bobot yang berbeda. Setiap topik (topik 0-4), mengandung kata kunci "aplikasi". Kata kunci "sertifikat" hanya terdapat pada topik 2 dan 3. Hal ini berarti bahwa topik berbeda dapat mengandung kata kunci yang sama, namun dengan nilai bobot yang berbeda pada setiap topiknya.

Setelah selesai membentuk topik, dilakukan pendistribusian topik ke data ulasan yang ada. Pendistribusian topik dilakukan untuk mengetahui setiap data ulasan memiliki kecenderungan untuk ditempatkan di antara topik 0 sampai dengan topik 4. Hasil keluarannya dapat dilihat pada Gambar 4. Dari hasil pendistribusian dapat dilihat bahwa hasilnya berupa tabel pengklasifikasian setiap data dokumen kedalam *dominant topic*, keywords pada setiap topik dan juga data teks dari dokumen 0 – 19999.

```
[(0,
  '0.057*"aplikasi" + 0.034*"masuk" + 0.025*"hp" + 0.022*"daftar" + 0.018*"pakai" + 0.018*"login" + 0.017*"nomor" +
0.015*"tolong" + 0.014*"coba" + 0.012*"buka"'),
 (1,
  '0.055*"aplikasi" + 0.029*"scan" + 0.023*"kode" + 0.023*"bagus" + 0.015*"tolong" + 0.015*"barcode" + 0.015*"update"
+ 0.014*"masuk" + 0.014*"buka" + 0.014*"error"'),
 (2,
  '0.074*"vaksin" + 0.067*"sertifikat" + 0.048*"aplikasi" + 0.026*"muncul" + 0.020*"data" + 0.015*"bantu" +
0.015*"peduli" + 0.014*"lindung" + 0.014*"cek" + 0.012*"mohon"'),
 (3,
  '0.055*"tanggal" + 0.040*"lahir" + 0.038*"isi" + 0.029*"data" + 0.021*"aplikasi" + 0.017*"nik" + 0.015*"sertifikat"
+ 0.014*"masuk" + 0.014*"sesuai" + 0.014*"tolong"'),
 (4,
  '0.028*"aplikasi" + 0.019*"buka" + 0.017*"lokasi" + 0.013*"update" + 0.011*"bikin" + 0.011*"izin" + 0.010*"tombol"
+ 0.009*"privasi" + 0.009*"susah" + 0.009*"habis"')]
```

Gambar 3. Hasil dari Pembuatan Topik

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	0.0	0.7026	vaksin, sertifikat, aplikasi, muncul, data, ba...	[mendinganagak, mudahterus, tingkat, karna, wa...
1	1	2.0	0.8224	tanggal, lahir, isi, data, aplikasi, nik, sert...	[klaim, sertifikat, data, isi, tombol, muncul,...
2	2	0.0	0.8852	vaksin, sertifikat, aplikasi, muncul, data, ba...	[vaksin, booster, sertifikat, aplikasi, peduli...
3	3	3.0	0.5942	aplikasi, scan, kode, bagus, tolong, barcode, ...	[fungsi]
4	4	3.0	0.6600	aplikasi, scan, kode, bagus, tolong, barcode, ...	[aplikasi, bagusternyata, update, biar, scan, ...
...	...	...	...	...	...
19995	19995	0.0	0.6346	vaksin, sertifikat, aplikasi, muncul, data, ba...	[aplikasi, kasih, bintang, moga, kembang, bala...
19996	19996	4.0	0.4772	aplikasi, buka, lokasi, update, bikin, izin, t...	[wifi, install, app, nyalain, bluetooth, jalan...
19997	19997	1.0	0.4328	aplikasi, masuk, hp, daftar, pakai, login, nom...	[kirim, otp, login, ngerti, aplikasi, kemarin,...
19998	19998	0.0	0.7316	vaksin, sertifikat, aplikasi, muncul, data, ba...	[daerah, sambut, aplikasi, sedia, versi, web, ...
19999	19999	0.0	0.5542	vaksin, sertifikat, aplikasi, muncul, data, ba...	[aplikasi, rating, bintang, maksud, bagus, val...

Gambar 4. Hasil dari Pendistribusian

Tabel 1. *Coherence Score*

Jumlah Topik	*Coherence Score*
1	0.3219
2	0.3481
3	0.3654
4	0.3899
5	0.3963
Rata-rata: 0.3643	

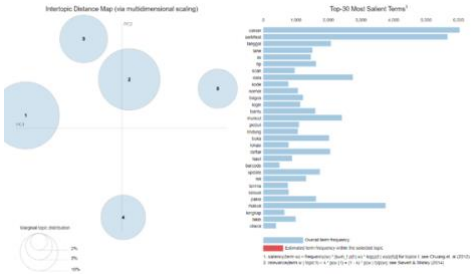
Evaluasi

Evaluasi *topic coherence* dilakukan setelah tahap pemodelan topik menggunakan metode LDA. Evaluasi ini dilakukan untuk mengetahui *coherence score* dari 5 topik yang telah dibentuk melalui pemodelan topik. Pada penelitian ini, jumlah topik akan dijadikan sebagai acuan dalam evaluasi. Nilai *coherence score* ditunjukkan pada Tabel 1. Pada Tabel 1. dapat dilihat dengan bahwa topik dengan jumlah sebanyak 5 menghasilkan *coherence score* sebesar 0.3963, dan *coherence score* secara keseluruhan memiliki nilai rata-rata sebesar 0.3643, maka dapat dikatakan bahwa topik dengan jumlah sebanyak 5 memiliki nilai yang cukup stabil.

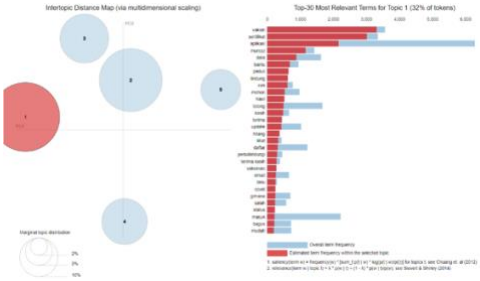
Visualisasi Data

Visualisasi data menampilkan hasil dari pemodelan topik ulasan pengguna PeduliLindungi. Visualisasi dihasilkan menggunakan *library gensim* pada pemrograman python yang berbentuk *Intertopic Distance Map*. Bentuk lingkaran yang berada di bagian kiri merupakan distribusi topik. Besar dan kecilnya diameter lingkaran menunjukkan persentase pada setiap topik. Jika diameter lingkaran tersebut besar, maka persentase topik tersebut tinggi dan berarti topik tersebut sering diulas. Jika diameter lingkaran tersebut kecil, maka persentase topik tersebut rendah dan berarti topik jarang diulas. Kualitas distribusi topik yang dihasilkan dipengaruhi apabila lingkaran saling tumpang tindih, apabila lingkaran tidak saling tumpang tindih maka hasilnya optimal, sedangkan apabila lingkaran saling tumpang tindih dan berkumpul di satu wilayah bagan, maka hasilnya kurang optimal. Posisi lingkaran yang terbentuk memiliki jarak kedekatan dengan lingkaran lainnya, kedekatan tersebut menunjukkan bahwa distribusi topik memiliki similaritas yang dekat berdasarkan model yang dibentuk.

Pemodelan topik divisualisasikan pada Gambar 5. Pada bagian kiri gambar ditampilkan 5 topik atau *Marginal Topic Distribution* berbentuk lingkaran, dan pada bagian kanan ditampilkan 30 kata yang paling sering muncul pada keseluruhan topik. Kata ‘vaksin’ merupakan kata yang paling sering muncul di dalam berbagai ulasan, kata ‘sertifikat’ merupakan kata terbanyak ke-dua yang sering muncul di dalam ulasan-ulasan, dan seterusnya. Terlihat bahwa kualitas distribusi topik optimal karena tidak tampak lingkaran yang tumpang tindih dan lingkaran tidak berkumpul di satu wilayah saja.



Gambar 5. Visualisasi Pemodelan Topik



Gambar 6. Visualisasi Pemodelan Topik 1

Distribusi topik 1 memiliki diameter lingkaran yang paling besar di antara semua topik yang terlihat. Hal tersebut dikarenakan topik 1 memiliki persentase sebesar 32% (Gambar 6) yang didapat dari perhitungan total token pada topik 1 sebesar 6400 data dibagi dengan total keseluruhan data yaitu 20.000, yang artinya topik 1 paling banyak dibicarakan di antara topik-topik yang lainnya. Lingkaran pada distribusi topik 1 juga tidak tumpang tindih dengan topik lain sehingga hasilnya dapat dikatakan optimal.

Persentase pada topik 2 sebesar 26.9%, yang artinya topik 2 cukup sering dibicarakan di bandingkan dengan topik 3, 4 dan 5. Persentase pada topik 3 sebesar 16.1%, yang artinya topik 3 jarang dibicarakan jika dibandingkan dengan topik 1 dan 2. Persentase pada topik 4 sebesar 14.1%, yang artinya topik 4 jarang dibicarakan di bandingkan dengan topik 1, 2 dan 3. Persentase pada topik 5 sebesar 10.9% yang artinya topik 5 sangat jarang dibicarakan jika dibandingkan dengan keempat topik lainnya.

KESIMPULAN DAN SARAN

Pemodelan topik pada ulasan pengguna aplikasi PeduliLindungi menggunakan metode *Latent Dirichlet Allocation* (LDA) menghasilkan sebanyak 5 topik. Jumlah topik yang ditentukan telah menghasilkan kumpulan kata yang membentuk topik dengan baik, dan memiliki persentase nya masing-masing.

Nilai *coherence score* dengan topik sebanyak 5 adalah 0,3963 dan dapat

disimpulkan bahwa topik sebanyak 5 memiliki hasil yang cukup baik. Informasi yang dihasilkan dari 5 topik yang dihasilkan pada penelitian ini dapat dimanfaatkan sebagai bahan evaluasi untuk meningkatkan pelayanan aplikasi PeduliLindungi. Persentase pada topik 1 sebesar 32% dari total keseluruhan data, yang artinya topik 1 paling banyak dibicarakan di antara topik-topik yang lainnya.

DAFTAR PUSTAKA

[1] A. Fastyaningsih, D. Priyantika, F. T. Widyastuti, K.Kismartini dan A. R. Herawati. "Keberhasilan Aplikasi Pedulilindungi Terhadap Kebijakan Percepatan Vaksinasi dan Akses Pelayanan Publik di Indonesia," *Jurnal Manajemen dan Kebijakan Publik*, vol. 6, no. 2, pp. 95-109, 2021.

[2] Menteri Komunikasi dan Informatika, *Keputusan Menteri Komunikasi dan Informatika Republik Indonesia Nomor 171*, 2020.

[3] Menteri Komunikasi dan Informatika, *Keputusan Menteri Komunikasi dan Informatika Republik Indonesia Nomor 253*, 2020.

[4] Kominfo, *Jumlah Pengguna PeduliLindungi Tembus 5% Pengguna Smartphone Indonesia*, 30 Juni 2020, https://www.kominfo.go.id.

[5] R. M. Snyder, "An Introduction to Topic Modeling as an Unsupervised Machine Learning Way to Organize Text Information," In Proc. ASCUE, 2015, pp. 86-96.

[6] P. Kherwa dan P. Bansal, "Topic Modeling: A Comprehensive Review," *ICST Transactions on Scalable Information Systems*, vol. 7, no. 24, 2018.

[7] T. Hua, C.-T. Lu, J. Choo, and C. K. Reddy, "Probabilistic Topic Modeling For Comparative Analysis of Document Collections," *ACM Transactions on Knowledge Discovery from Data*, vol. 14, no. 2, pp. 1-27, 2020.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] M. A. Nugraha, dan L. C. Mungaran, "Pemodelan Topik Berita pada Portal Berita Online Berbahasa Indonesia Menggunakan Latent Dirichlet Allocation (LDA)," *Jurnal Ilmiah Komputasi*, vol. 20, no. 2, pp. 173-180, 2021.

[10] D. Purwitasari, A. Muflichah, N. A. Hasanah, dan A. Z. Arifin, "Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir," Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI), vol. 5, no. 3, pp. 421–428, 2021. [Online serial]. Available: https://doi.org/10.29207/resti.v5i3.3049 . [Accessed March 17, 2022].

[11] A. I. Alfanzar, Khalid, dan I. S. Rozas,

"Topic Modelling Skripsi Menggunakan Metode Latent," *Jurnal Sistem Informasi (JSiI)*, vol. 7, no. 1, pp. 7–13, 2020.

[12] E. P. A. Akhmad, dan C. L. Prawirosastro, "Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online," Laporan Hasil Penelitian, Surabaya: KPN, Universitas Hang Tuah, 2021.

[13] Patmawati, dan M. Yusuf. "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 3, pp. 122–129, 2021.

[14] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, 2010.

[15] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocessing Techniques for Text Mining - An Overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7-16, 2015.

[16] V. Gupta, and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp.60-76, 2009.