

Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA)

Uray Nur Khadijah¹, Nuri Cahyono^{2*}

uraynurkhadijah@students.amikom.ac.id¹, nuricahyono@amikom.ac.id²

^{1,2} Universitas Amikom Yogyakarta

Informasi Artikel

Diterima : 8 Mar 2024

Direvisi : 11 Jun 2024

Disetujui : 25 Jul 2024

Kata Kunci

Yogyakarta, social media, Platform, Twitter, Topic Modelling, Latent Dirichlet Allocation (LDA)

Abstrak

Pariwisata Yogyakarta sebagai destinasi yang kaya akan budaya dan sejarah, sering menjadi fokus diskusi di media sosial. Tujuan dari Penelitian ini adalah menelaah topik pariwisata Yogyakarta dari *Twitter*. Dataset yang diperoleh dalam penelitian ini dari *crawling* data menggunakan *API key Twitter*. Penelitian ini menggunakan tahapan dari pengumpulan data, text preprocessing, dan menerapkan metode *Topic Modelling*, khususnya *Latent Dirichlet Allocation (LDA)*. Hasil penelitian ini pengujian kinerja pemodelan topik dengan metode LDA dapat dilihat dari nilai *coherence score*, semakin tinggi nilai *coherence* suatu topik, semakin mudah diinterpretasikan oleh manusia dan *Perplexity* merupakan salah satu standar pengukuran yang dapat digunakan untuk menilai kinerja model yang baik dari model tersebut ditunjukkan dengan nilai *perplexity* yang lebih rendah. Nilai *coherence score* yang ditunjukkan pada num topic ke-1 sebesar 0.331047, untuk nilai *perplexity* ditunjukkan dengan nilai yang tinggi terletak pada num topic ke-3 sebesar -8.830172565520245 diharapkan dapat memberikan wawasan mendalam tentang topik-topik yang sering dibahas dan berkonsentrasi pada penerapan sistem pemodelan topik untuk membangun sistem keputusan topik berita yang menggunakan metode Latent Dirichlet Allocation (LDA). Pada Penelitian ini efektif dalam menggunakan metode LDA untuk menentukan topik berita yang mencakup tiga kategori topik yang sering dibicarakan pada masing-masing kelas.

Keywords

Yogyakarta, social media, Platform, Twitter, Topic Modeling, Latent Dirichlet Allocation (LDA), crawling, coherence score, perplexity.

Abstract

Yogyakarta tourism, as a destination rich in culture and history, is often the focus of discussion on social media. The purpose of this research is to examine the topic of Yogyakarta tourism from Twitter. The dataset obtained in this research is from crawling data using the Twitter API key. This research uses the stages of data collection, text preprocessing, and applying the Topic Modeling method, specifically Latent Dirichlet Allocation (LDA). The results of this study test the performance of topic modeling with the LDA method can be seen from the coherence score value, the higher the coherence value of a topic, the easier it is to be interpreted by humans and Perplexity is one of the measurement standards that can be used to assess the performance of a good model of the model indicated by a lower perplexity value. The coherence score value shown in the 1st num topic is 0.331047, for the perplexity value shown with a high value located in the 3rd num topic of -8.830172565520245. is expected to provide deep insight into topics that are often discussed and concentrate on the application of topic modeling systems to build news topic decision systems that use the Latent Dirichlet Allocation (LDA) method. This research is effective in using the LDA method to determine news topics that include three categories of frequently discussed topics in each class.

A. Pendahuluan

Pariwisata merupakan salah satu sector penting dalam perekonomian Indonesia, termasuk Yogyakarta. Yogyakarta terkenal dengan kekayaan budaya, sejarah, dan alamnya yang menarik banyak wisatawan. Untuk meningkatkan daya Tarik wisata Yogyakarta dan apa yang mereka cari dalam pengalaman wisata mereka[1]. Media sosial, khususnya Twitter, telah menjadi platform yang populer bagi wisatawan untuk berbagi pengalaman dan pendapat mereka tentang destinasi wisata. Data tweet dapat menjadi sumber informasi yang berharga untuk memahami persepsi wisatawan terhadap Yogyakarta. Dan platform ini, pengguna Twitter berbagi pemikiran, rekomendasi, serta pengalaman mereka terkait perjalanan dan kunjungan ke Yogyakarta[2].

Penggunaan metode analisis teks, khususnya Topic Modeling, dapat membuka potensi data Twitter untuk mengidentifikasi topik-topik utama yang muncul dalam percakapan seputar pariwisata Yogyakarta. Salah satu metode yang populer dalam analisis topik adalah Latent Dirichlet Allocation (LDA), yang memungkinkan identifikasi pola tersembunyi dalam dataset teks[3].

Penelitian-penelitian sebelumnya telah menunjukkan keberhasilan LDA dalam berbagai domain, termasuk analisis data dari media sosial. Penerapan metode ini pada dataset Twitter terkait pariwisata Yogyakarta dapat memberikan pemahaman lebih dalam topik-topik utama yang muncul dalam percakapan di platform ini. Dengan demikian, penelitian ini bertujuan untuk menggali wawasan baru dengan menerapkan LDA pada dataset Twitter terkait pariwisata Yogyakarta. Melalui analisis ini, diharapkan akan ditemukan topik-topik yang domain dan dinamika unik dari percakapan media sosial terkait destinasi ini[4].

B. Metode Penelitian

Metode yang digunakan dalam penelitian mulai dari pengambilan data, pembersihan data (preprocessing) dan interpretasi hasil penelitian sehingga dapat menjadi sesuatu hal yang bermanfaat.

1. Pengambilan Data

Pengambilan data dilakukan dengan menggunakan crawling data menggunakan google collab, yang dalam mengambil data dari twitter dengan rentang waktu September 2021 – januari 2024. Crawling data Twitter adalah proses pengambilan data dari Twitter secara otomatis dengan menggunakan program computer Bahasa python. Proses ini dilakukan dengan memanfaatkan API twitter yang menyediakan akses ke data tweet, seperti teks, profil pengguna, dan lainnya. Dan hasil dari crawling data ini mendapatkan 7758 dataset. Data yang dikumpulkan untuk penelitian ini berasal dari komentar masyarakat (netizen) tentang pariwisata di yogyakarta dari cuitan Twitter.

2. Text Processing

Preprocessing data adalah Langkah selanjutnya, yang menghapus beberapa symbol dari analisis topik dan tahap pertama dalam mulai menganalisis. Ini penting karena membuat teks tidak berstruktur dan menyimpan kata kunci yang membantu menyajikan topik. Pada tahap pengolahan teks, tahap ini sangat penting untuk menghilangkan kata-kata yang tidak diperlukan dari dokumen. Berikut

adalah beberapa Langkah yang digunakan dalam penelitian ini untuk preprocessig data :

a. Case folding

Proses ini untuk mengubah seluruh karakter pada tweet menjadi huruf kecil.

b. Tokenizing

Tahapan *tokenizing* merupakan proses analisis untuk memenggal kata yang Menyusun suatu kalimat. *Tokenizing* mencakup proses *removing number*, *removing punctuation* seperti *symbol* dan tanda baca yang tidak penting,

c. Normalization

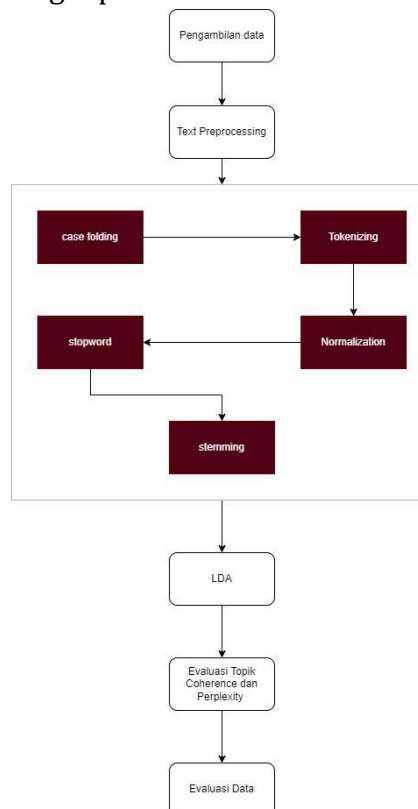
Proses *normalization* adalah langkah untuk membawa teks ke dalam bentuk standar atau normal.

d. Stopword

Stopword dilakukan untuk membersihkan teks dari kata-kata yang umum dan tidak relevan ataupun tidak memiliki arti.

e. Stemming

Tahap stemming ini untuk menghapus sufiks didalam suatu.



Gambar 1. Alur Penelitian

3. *Latent Dirichlet Allocation (LDA)*

Latent Dirichlet Allocation (LDA) adalah sebuah model *statistic generative* yang digunakan untuk mengidentifikasi pola tematik yang tersembunyi dalam koleksi dokumen. Model ini didasarkan pada asumsi bahwa setiap dokumen terdiri dari campuran topik, dan setiap topik diwakili oleh distribusi kata-kata dan memiliki beberapa manfaat, seperti analisis topik, klasifikasi dokumen, dan pengurangan dimensi. Namun, LDA juga memiliki beberapa kekurangan, seperti asumsi, interpretasi, dan kompleksitas. Cara kerja metode LDA adalah dengan mengambil topik yang telah ditetapkan sebelum mendapatkan dokumen, kemudian menerapkan proses pada setiap dokumen dalam koleksi.

4. *Perplexity*

Dengan menggunakan evaluasi model, penelitian ini akan menentukan jumlah topik yang ideal tanpa pelatihan data. *Perplexity* adalah metode yang paling umum untuk menilai model LDA. *Perplexity* adalah metrik yang digunakan untuk mengevaluasi ketepatan informasi dokumen tentang topik yang dihasilkan. Nilai *perplexity* dihitung dengan menghitung kemungkinan dari log teks dokumen yang tidak terlihat; nilai yang lebih rendah menunjukkan model yang lebih baik dan menunjukkan lebih sedikit ketidakpastian tentang dokumen yang tidak terlihat.

5. *Coherence*

Coherence dalam konteks topic modeling, khususnya *Latent Dirichlet Allocation (LDA)*, mengacu pada pengukuran sejauh mana topik yang dihasilkan oleh model memiliki keterkaitan atau koherensi. Koherensi digunakan untuk mengevaluasi sejauh mana topik-topik yang dihasilkan oleh model benar-benar mencerminkan struktur semantik dalam dataset. Dalam konteks evaluasi LDA, ukuran koherensi dapat membantu menentukan apakah topik-topik yang dihasilkan oleh model dapat dianggap bermakna dan konsisten. Terdapat beberapa metode untuk mengukur koherensi, dan umumnya digunakan metode seperti uji koherensi *c_v*, uji koherensi *u_mass*, atau metode lainnya. Skor koherensi yang lebih tinggi menunjukkan bahwa topik-topik yang dihasilkan memiliki hubungan semantik yang lebih kuat di antara kata-kata yang membentuk topik tersebut. Evaluasi koherensi penting untuk memastikan bahwa model LDA dapat menghasilkan topik-topik yang relevan dan konsisten dengan struktur informasi dalam dataset. Sebuah model dengan koherensi yang tinggi cenderung lebih dapat dipercaya dalam menjelaskan struktur topikal dalam data teks.

C. Hasil dan Pembahasan

Dalam penelitian ini pengolahan data dilakukan dengan tahap preprocessing dengan Metode pembagian Latent Dirichlet (LDA). Tujuan metode ini adalah untuk mengidentifikasi kata topik yang mewakili setiap topik. Metode ini digunakan dengan library python, dan hasil pengolahan data yang berupa diagram yang mana didalamnya mencakup 3 topik berita yang sering dibahas di akun media sosial Twitter tentang pariwisata yogyakarta serta melakukan perhitungan nilai dari topic coherence dan perplexity, untuk topik coherence yang dimana menunjukkan semakin tinggi nilai coherence maka model yang dihasilkan

akan semakin baik sedangkan untuk nilai perplexity, semakin kecil nilai perplexity yang dihasilkan menunjukan model yang lebih baik.

1. Visualisasi data LDA

Hasil visualisasi data *TopicModelling* ini, Dengan menggunakan library PyLDAvis untuk visualisasi dan library python dalam implementasi, hasil pemodelan topik ini ditampilkan secara visual. Library ini membantu dalam penafsiran topik yang sesuai dengan dokumen, dan mengekstraksi informasi dari model LDA yang dibuat dengan memberikan informasi visual dalam bentuk tampilan web yang interaktif. Tampilan visual ini dapat digunakan dalam *Google Collab*. Hasil pengolahan data ditampilkan dalam bentuk diagram yang menggambarkan isi berita yang mencakup 3 topik yang paling sering dibahas di media sosial Twitter tentang pariwisata yogyakarta.

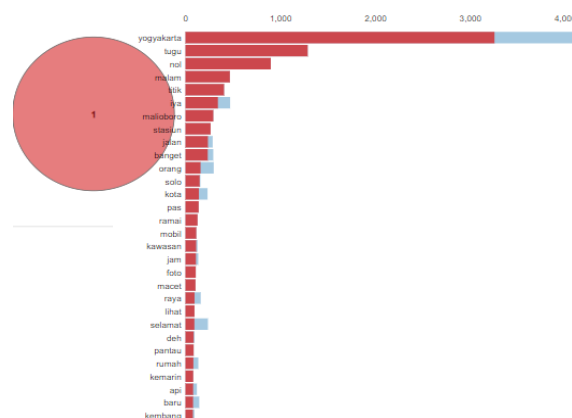
A. Topik 1

Bagian ini dibagi menjadi 3 topik untuk 1 topik utamanya terdapat 30 pembahasan dalam satu topik. Topik pertama lebih membahas mengenai kehidupan sehari-hari, lalu lintas, dan aktivitas di kota Yogyakarta serta respons atau perasaan terhadap pengalaman. Representasi topik-topik yang ditemukan oleh model LDA berdasarkan matriks dokumen-istilah yang telah dibuat. Berikut pada tabel 1.

Tabel 1. Model LDA Topic 1

Model LDA Topik 1
$[(0.038 \cdot \text{"yogyakarta"} + 0.008 \cdot \text{"libur"} + 0.006 \cdot \text{"selamat"} + 0.006 \cdot \text{"iya"} + 0.006 \cdot \text{"wisata"} + 0.004 \cdot \text{"informasi"} + 0.004 \cdot \text{"kerja"} + 0.004 \cdot \text{"kota"} + 0.004 \cdot \text{"diy"} + 0.004 \cdot \text{"cek"}),$

Pada tabel 1 tersebut merupakan representasi dari topik-topik yang dihasilkan oleh model LDA. Setiap tuple pada output menyajikan informasi mengenai suatu topik, dengan kata-kata kunci dan bobotnya. Di bawah ini, Gambar 2 menunjukkan bahwa jika salah satu topik dipilih, lingkaran akan berubah warna menjadi merah. Selanjutnya, bar chart di sisi kanan panel akan berubah warna merah,



menunjukkan estimasi term frekuensi topik yang dipilih.

Gambar 2. Diagram Topic 1

Selanjutnya, model LDA dapat ditampilkan dalam bentuk wordcloud seperti gambar 5.



Gambar 5 Wordcloud Topik 2

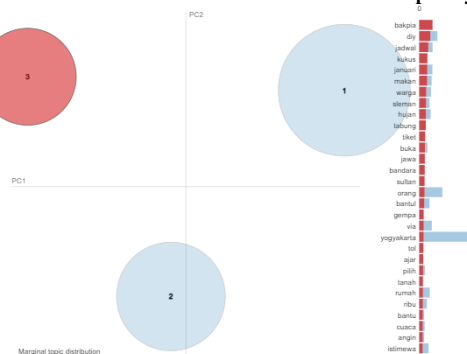
C. Topik 3

Topik akhir dari penelitian ini yaitu topik 3, yang memuat informasi seputan beragam aspek kehidupan sehari-hari, keadaan alam, infrastruktur, dan tempat-tempat penting di Yogyakarta dan sekitarnya. Representasi topik-topik yang ditemukan oleh model LDA berdasarkan matriks dokumen-istilah yang telah dibuat. Berikut pada tabel 3.

Tabel 3. Model LDA Topic 2

Model LDA Topic 3
(2, '0.100*"yogyakarta" + 0.040*"tugu" + 0.028*"nol" + 0.014*"malam" + 0.013*"titik" + 0.011*"iya" + 0.009*"malioboro" + 0.008*"stasiun" + 0.007*"jalan" + 0.007*"banget""])

Pada tabel 4.10 tersebut merupakan representasi dari topik-topik yang dihasilkan oleh model LDA. Setiap tuple pada output menyajikan informasi mengenai suatu topik, dengan kata-kata kunci dan bobotnya. Di bawah ini, Gambar 4.7 menunjukkan bahwa jika salah satu topik dipilih lingkaran akan berubah warna menjadi merah. Selanjutnya, bar chart di sisi kanan panel akan berubah warna merah, menunjukkan estimasi term frekuensi topik yang dipilih.



Gambar 6. Diagram Topik 3

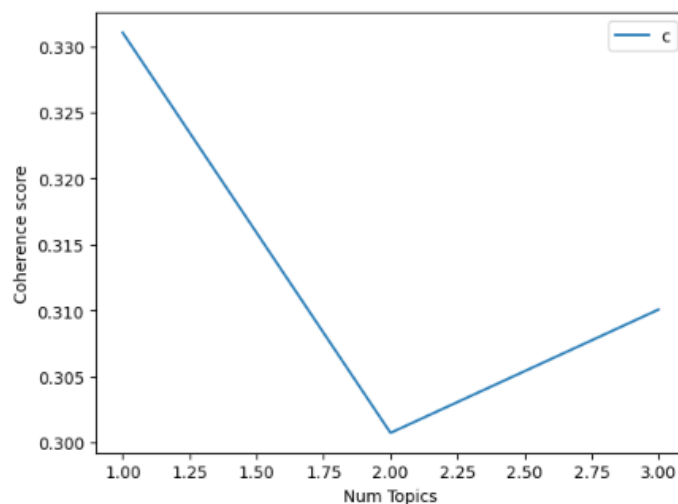
Selanjutnya, model LDA dapat ditampilkan dalam bentuk wordcloud seperti gambar 6.



Gambar 6. Wordcloud Topik 3

2. Topic coherence

Topic coherence merupakan tahapan untuk menentukan jumlah model topik dalam suatu topic modelling LDA. Nilai yang lebih tinggi dari topic coherence menunjukkan bahwa model yang dihasilkan akan lebih baik. Jumlah topik dengan nilai coherence yang lebih tinggi akan menunjukkan bahwa model yang dihasilkan akan lebih baik. Tertinggi akan digunakan sebagai topic modelling. Grafik berikut menunjukkan hasil dari subjek koherensi.



Gambar 7. Topic Coherence

Gambar 7 menunjukkan bahwa titik tertinggi dalam grafik koherensi topik berada pada jumlah topik ke 1. Untuk memudahkan melihat hasil visualisasi dari gambar 7 seperti ditunjukkan dengan menggunakan nilai coherence score yang ditunjukkan dalam tabel 4 dibawah ini.

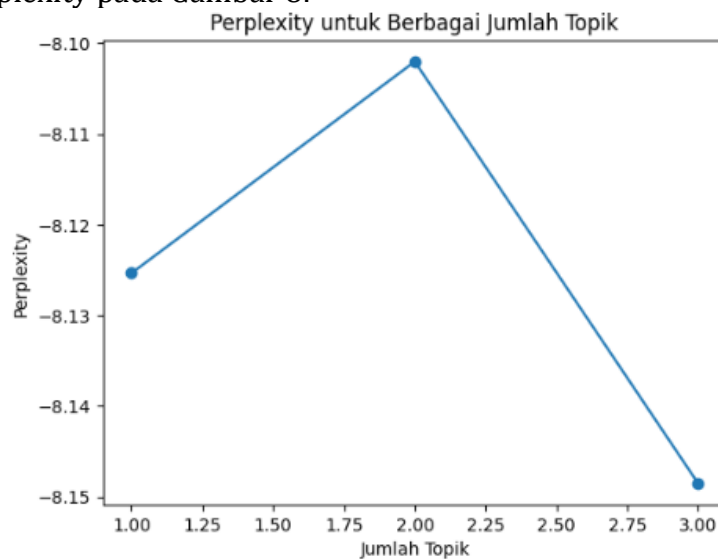
Tabel 4. Nilai Topic Coherence

Num topics	Coherence score
1	0.331047
2	0.300728
3	0.310066

Pada tabel diatas menunjukkan hasil nilai dari topic coherence yang tertinggi terletak pada *num topic* ke-1 sebesar 0.331047.

3. Perplexity

Perplexity dapat digunakan untuk membandingkan berbagai jenis pemodelan bahasa, tetapi perbandingan ini memerlukan pemodelan bahasa dan teks yang akan diujikan. Menggunakan teks yang diujikan yang sama. Karena pengurangan kardinalitas dapat menurunkan jumlah kata yang mungkin, ukuran kosa kata dapat dengan mudah dikaitkan dengan perplexitas[17]. Dapat dilihat visualisasi perplexity pada Gambar 8.

**Gambar 8.** Perplexity

Dari hasil visualisasi pada gambar 8 dapat dilihat di topik 3 memiliki nilai yang rendah artinya model lebih baik. Untuk memudahkan melihat hasil visualisasi dari gambar 8 seperti ditunjukkan dengan menggunakan nilai Perplexity yang ditunjukkan dalam Tabel 4.12 dibawah ini.

Tabel 5. Nilai Perplexity

Num topics	Perplexity
1	-8.811879362463792
2	-8.805022364846005
3	-8.830172565520245

F. Referensi

- [1] N. L. P. M. Putu, Ahmad Zuli 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, 2021, doi: 10.29207/resti.v5i1.2587.
- [2] V. R. Hananto, U. Dinamika, and L. D. Allocation, "Analisis Topik pada Forum Pariwisata di Indonesia Menggunakan Latent Dirichlet Allocation," vol. 17, no. x, pp. 253–262, 1978.
- [3] F. Rashif, G. Ihza Perwira Nirvana, M. Alif Noor, and N. Aini Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19," *Cogito Smart J.*, vol. 7, no. 1, pp. 1–12, 2021.
- [4] "Indonesia Menggunakan Metode Online Latent Dirichlet Allocation," 2022.
- [5] B. Griciūtė, L. Han, and G. Nenadic, "Topic Modelling of Swedish Newspaper Articles about Coronavirus: a Case Study using Latent Dirichlet Allocation Method," no. M1, 2023, [Online]. Available: <http://arxiv.org/abs/2301.03029>
- [6] N. Cahyono and Angga Reni Dwi Astuti, "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.
- [7] D. F. O. Onah, E. L. L. Pang, and M. El-Haj, "A Data-driven Latent Semantic Analysis for Automatic Text Summarization using LDA Topic Modelling," *Proc. - 2022 IEEE Int. Conf. Big Data, Big Data 2022*, pp. 2771–2780, 2022, doi: 10.1109/BigData55660.2022.10020259.
- [8] F. Ahmed, M. Nawaz, and A. Jadoon, "Topic Modeling of the Pakistani Economy in English Newspapers via Latent Dirichlet Allocation (LDA)," *SAGE Open*, vol. 12, no. 1, 2022, doi: 10.1177/21582440221079931.
- [9] I. Kharisudin and H. Masri'an, "Topic Modeling on WhatsApp User Reviews Using Latent Dirichlet Allocation," *Sci. J. Informatics*, vol. 9, no. 1, pp. 51–62, 2022, doi: 10.15294/sji.v9i1.34941.
- [10] G. Rosalinda, R. Santoso, and P. Kartikasari, "Pemodelan Topik Ulasan Aplikasi Netflix Pada Google Play Store Menggunakan Latent Dirichlet Allocation," *J. Gaussian*, vol. 11, no. 4, pp. 554–561, 2023, doi: 10.14710/j.gauss.11.4.554-561.
- [11] K. H. Musliadi, H. Zainuddin, and Y. Wabula, "Twitter Social Media Conversion Topic Trending Analysis Using Latent Dirichlet Allocation Algorithm," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 390–399, 2022, doi: 10.37385/jaets.v4i1.1143.
- [12] A. Kim, S. Sevanto, E. R. Moore, and N. Lubbers, "Latent Dirichlet Allocation modeling of environmental microbiomes," *PLoS Comput. Biol.*, vol. 19, no. 6, pp. 1–30, 2023, doi: 10.1371/journal.pcbi.1011075.

- [13] W. A. Nawang Sari and H. Dwi Purnomo, "Topic Modeling Using the Latent Dirichlet Allocation Method on Wikipedia Pandemic Covid-19 Data in Indonesia," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1223–1230, 2022, doi: 10.20884/1.jutif.2022.3.5.321.
- [14] M. O. Ajinaja, A. O. Adetunmbi, C. C. Ugwu, and O. S. Popoola, "Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling," *Iran J. Comput. Sci.*, vol. 6, no. 1, pp. 81–94, 2023, doi: 10.1007/s42044-022-00124-7.
- [15] 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.
- [16] D. Oleh, U. H. Tuah, P. D. Pelayaran, and U. H. Tuah, "Hasil_Peneltian_21_Ekka-Carlos," 2021.
- [17] T. Santika, *Evaluasi Perplexity Untuk Pemodelan Topik Diskusi Agama Islam di Media Sosial Twitter Indonesia Tahun 2006-2018 Menggunakan Latent Dirichelet Allocation*. 2019.
- [18] M. Luvian chisni chilmi, "Latent dirichlet allocation (LDA) untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law," *Univ. Islam Negeri Syarif Hidayatullah*, pp. 1–131, 2021, [Online]. Available: <https://repository.uinjkt.ac.id/dspace/handle/123456789/56724>