

# IMPLEMENTASI ALGORITMA *LATTERN DIRICHLET ALLOCATION (LDA)* DENGAN *GIBSS SAMPLING* UNTUK *TOPIC MODELLING* MENGGUNAKAN DATA X TERKAIT IKN

Raihan Labib Hanif<sup>1</sup>, Mohammad Syafrullah<sup>2\*</sup>

<sup>1,2\*</sup> Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta  
Email: <sup>1</sup>2011501315@student.budiluhur.ac.id, <sup>2\*</sup>mohammad.syafrullah@budiluhur.ac.id

(Naskah masuk: 7 Agustus 2024, diterima untuk diterbitkan: 7 September 2024)

## Abstrak

Penelitian ini membahas tentang pemodelan topik terkait tweet masyarakat mengenai Ibu Kota Negara (IKN) di media sosial X. Masalah utama yang diangkat adalah kurangnya pemahaman mengenai tren topik yang sering dibahas masyarakat terkait IKN, yang dapat mempengaruhi persepsi dan opini publik. Untuk menyelesaikan masalah ini, metode Latent Dirichlet Allocation (LDA) digunakan dalam pemodelan topik, yang memungkinkan pengelompokan topik secara otomatis berdasarkan teks yang tersedia. Proses pemodelan dilakukan dengan menggunakan teknik Gibbs Sampling dan diukur menggunakan nilai coherence untuk mengevaluasi kualitas model yang dihasilkan. Tujuan dari penelitian ini adalah untuk mengidentifikasi dan mengelompokkan topik yang muncul dari tweet terkait IKN, sehingga dapat memberikan wawasan yang lebih mendalam mengenai pandangan masyarakat. Hasil penelitian menunjukkan bahwa model dengan 20 topik menghasilkan nilai coherence tertinggi sebesar 0.38389228712614887, yang menunjukkan bahwa model ini efektif dalam mengidentifikasi tren topik yang dibahas oleh masyarakat.

**Kata kunci:** pemodelan topik, gibss sampling, coherence score, IKN, latent dirichlet allocation (LDA)

## IMPLEMENTATION OF *LATTERN DIRICHLET ALLOCATION (LDA)* ALGORITHM WITH *GIBSS SAMPLING* FOR *TOPIC MODELLING* USING X DATA RELATED TO IKN

### Abstract

This research discusses topic modeling related to public tweets regarding the National Capital City (IKN) on X social media. The main problem raised is the lack of understanding of topic trends that are often discussed by the public related to the National Capital, which can affect public perception and opinion. To solve this problem, the Latent Dirichlet Allocation (LDA) method is used in topic modeling, which allows automatic clustering of topics based on available text. The modeling process is performed using the Gibbs Sampling technique and measured using the coherence value to evaluate the quality of the resulting model. The purpose of this research is to identify and categorize topics that arise from tweets related to IKN, so as to provide a deeper insight into the views of the public. The results show that the model with 20 topics produces the highest coherence value of 0.38389228712614887, which indicates that the model is effective in identifying trends in topics discussed by the public.

**Keywords:** topic modelling, gibss sampling, coherence score, IKN, latent dirichlet allocation (LDA)

## 1. PENDAHULUAN

Teknologi informasi mencakup apapun yang membantu manusia membuat, mengubah, menyimpan, mengomunikasikan, atau menyebarkan informasi. Tanpa teknologi ini, hubungan sosial tidak dapat terbentuk. Saat ini, perantara media cetak tidak lagi digunakan untuk menyampaikan informasi, tetapi diganti dengan media digital elektronik [1]. Fokus utama penelitian ini adalah teknologi teks yang dapat digunakan untuk mendapatkan topik dari kumpulan kalimat [2]. Tidak terlepas dari peran manusia yang terus menerus menciptakan dan

menyempurnakan teknologi, kemajuan teknologi terus berkembang dengan pesat [3].

Media sosial adalah platform yang memungkinkan individu untuk berbagi gagasan, pandangan, dan informasi melalui jaringan dan komunitas digital. X merupakan salah satu media sosial yang paling populer dan banyak digunakan oleh publik, yang menawarkan layanan jaringan sosial di mana pengguna dapat berbagi dan berinteraksi dengan pesan singkat yang dikenal sebagai "tweet" [4]. X Inc. mengoperasikan dan mengelola platform jejaring sosial berbasis microblogging yang dikenal sebagai X, yang

memungkinkan pengguna untuk mengirim dan membaca pesan singkat ke sesama pengguna [5].

IKN adalah singkatan dari Ibu Kota Negara. Dalam konteks Indonesia, IKN mengacu pada rencana pemindahan ibu kota pemerintahan dari Jakarta ke lokasi baru di Kalimantan Timur. Inisiatif ini diajukan oleh Presiden Jokowi dan disetujui oleh DPR RI melalui rapat paripurna pada 18 Januari 2022. Pemindahan ibu kota ini ke wilayah Sepaku, Penajam Paser Utara, diprediksi akan membawa dampak yang signifikan pada aspek ekonomi, sosial, dan budaya [6].

Pembahasan orang terkait IKN pada media sosial X cukup beragam, sehingga dalam penelitian ini merasa diperlukan adanya pemodelan topik untuk mengetahui topik apa yang sering dibahas oleh masyarakat.

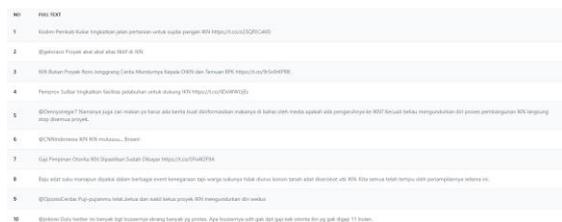
Adapun penelitian terdahulu [7], Metode yang digunakan dalam penelitian ini adalah *Latent Dirichlet Allocation (LDA)*, yang berfungsi sebagai pemodelan topik untuk data *tweet* tentang tokoh publik politik. Penelitian ini memanfaatkan metode *Agglomerative Hierarchical Clustering* untuk mengelompokkan data *tweet* berdasarkan kemiripan, sehingga menghasilkan beberapa kluster topik. Setelah proses klustering, metode LDA diterapkan untuk mengidentifikasi distribusi topik dalam dokumen, dengan parameter yang telah diinisialisasi seperti jumlah topik, nilai  $\alpha$  (*Dirichlet*), nilai  $\beta$ , dan jumlah iterasi. Hasil dari LDA kemudian diolah lebih lanjut menggunakan *TF-IDF* untuk mengetahui *tweet* mana saja yang memiliki kata-kata terbanyak dalam topik yang dihasilkan, dan menggunakan metode *Silhouette Coefficient* dalam evaluasi hasil klustering

## 2. METODE PENELITIAN

Dalam penelitian ini dilakukan pengambilan data dalam bentuk teks, berupa data *tweet* dari tanggal 1 Januari 2024 hingga 11 Juli 2024 yang diambil melalui situs resmi yaitu X, dengan total data sebanyak 216 *tweet* yang membahas tentang IKN.

### 2.1 Data Penelitian

Pada tahap pengumpulan data ini menggunakan *tweet harvest*. Pada tahapan *crawling*, kata kunci atau *keyword* yang digunakan berkaitan dengan IKN. Data yang diambil dari proses *crawling* akan disimpan dalam bentuk file CSV. Kemudian, data yang sudah menjadi format CSV akan disimpan pada *database* lokal untuk dijadikan dataset.



Gambar 1. Data Kotor

### 2.2 Preprocessing

*Preprocessing* adalah tahap awal dalam menyiapkan dataset untuk mempermudah pemrosesan data. Tujuannya adalah untuk meningkatkan kualitas data sehingga menghasilkan performa yang tinggi [8]. Tahap-tahap *preprocessing* yang digunakan antara lain sebagai berikut:

#### a) Stopword Removal

Proses penghapusan *stopword* adalah langkah untuk menghilangkan kata-kata penghubung dan kata-kata yang tidak relevan dalam dataset tersebut [9]. Langkah ini akan menghilangkan semua kata yang dianggap tidak memiliki makna signifikan, seperti kata penghubung [10].

#### b) Stemming

Stemming merupakan metode yang digunakan untuk menemukan bentuk dasar dari kata yang memiliki imbuhan dengan cara menghilangkan imbuhan tersebut. Proses stemming terdiri dari tiga langkah utama: pertama, mengidentifikasi dan mengumpulkan aturan-aturan yang terdapat dalam bahasa Balantak; kedua, mengumpulkan kata dasar yang akan disimpan dalam kamus data; dan ketiga, menghapus imbuhan yang ada pada kata dasar [11].

#### c) Case folding

*Case folding* adalah tahap proses yang mengubah kata menjadi bentuk seragam menggunakan metode lower pada string di Python. Tujuan dari *Case folding* adalah mengubah semua kata menjadi huruf kecil agar teks yang diproses berada dalam format yang konsisten. *Cleaning* teks adalah proses membersihkan dan memvalidasi kata-kata yang tidak diinginkan untuk mengurangi gangguan dalam proses klasifikasi. Contohnya, menghilangkan karakter tertentu [12].

### 2.3 Topic Modelling

Topik modeling adalah metode dalam text mining yang digunakan untuk menemukan pola dalam dokumen serta mengelompokkan kata-kata ke dalam beberapa topik yang mencerminkan konten dari dokumen tersebut. Teknik ini menggunakan pendekatan pengelompokan di mana kata-kata dalam teks dikategorikan menjadi topik-topik tertentu. Latent Dirichlet Allocation (LDA) merupakan salah satu teknik yang sering digunakan dalam topik modeling untuk mengekstrak topik-topik penting dari dokumen. LDA dikenal efektif dalam menangani data yang berukuran besar [13].

### 2.4 LDA dengan Gibbs Sampling

Mengaplikasikan LDA dengan *Gibbs Sampling* melibatkan kalkulasi variabel yang dibutuhkan, inialisasi acak untuk topik, serta menjalankan sejumlah iterasi yang telah ditentukan. Pada setiap iterasi, topik akan disampel untuk setiap kata dalam *corpus*. Setelah semua iterasi selesai, hasilnya digunakan untuk menghitung distribusi laten  $\theta_d$  dan  $\phi_k$ . Adapun variabel-variabel yang dibutuhkan adalah sebagai berikut:

$n_{d_i k}$  = jumlah dokumen  $d$  pada topik  $k$ .  
 $n_{k,w}$  = jumlah kata  $w$  yang diberikan ke topik  $k$ .  
 $n_k$  = jumlah total dari setiap kata yang diberikan ke topik  $k$ .  
 $n_d$  = jumlah total seluruh dokumen  $d$  dalam semua topik.

Karena prosedur pengambilan sampel *Gibbs* melibatkan pengambilan sampel dari distribusi yang dikondisikan pada semua variabel, kita harus menghapus penugasan topik saat ini sebelum membangun distribusi dari persamaan:

$$p(z_i = k | w_i, d_i, \alpha, \beta) \propto \frac{n_{d_i k} + \alpha}{\sum_{k'=1}^K n_{d_i k'} + K\alpha} \cdot \frac{n_{k w_i} + \beta}{\sum_{w=1}^W n_{k w} + W\beta} \dots (1)$$

Keterangan:

$z_i = k$  = pemberian topik  $k$  untuk kata ke- $i$   
 $w_i$  = kata ke- $i$   
 $d_i$  = dokumen ke- $i$   
 $\alpha$  = *hyperparameter* untuk prior *Dirichlet* pada distribusi topik-dokumen  
 $\beta$  = *hyperparameter* untuk prior *Dirichlet* pada distribusi kata-topik  
 $n_{d_i k}$  = jumlah dokumen  $d$  pada topik  $k$   
 $\sum_{k'=1}^K n_{d_i k'}$  = total seluruh dokumen  $d$  dalam semua topik  $k$   
 $n_{k w_i}$  = jumlah kata  $w$  yang muncul pada topik  $k$   
 $\sum_{w=1}^W n_{k w}$  = jumlah kata pada topik  $k$   
 $K$  = jumlah topik  
 $W$  = panjang dari kamus kata

Penghapusan dilakukan dengan mengurangi jumlah yang terkait dengan penugasan saat ini, karena urutan penugasan topik tidak mempengaruhi hasil akhir. Selanjutnya, probabilitas dari setiap penugasan topik dihitung menggunakan persamaan yang relevan. Distribusi ini kemudian diambil sampelnya, dan topik yang dipilih diatur dalam array  $z$ , serta jumlah yang sesuai ditambahkan kembali.

## 2.5 Nilai Coherence

*Topic Coherence* adalah metrik yang digunakan untuk menilai suatu topik dengan mengukur tingkat kesamaan semantik antara kata-kata dalam topik tersebut. Nilai *topic coherence* yang lebih tinggi menunjukkan bahwa topik lebih mudah diinterpretasikan oleh manusia, sehingga menghasilkan kualitas topik yang lebih baik [13].

*Coherence score* adalah ukuran yang digunakan untuk mengevaluasi pemodelan topik. Model yang baik akan menghasilkan topik dengan skor koherensi yang tinggi, yang menunjukkan bahwa topik-topik tersebut saling terkait dan memiliki keterkaitan yang logis dalam kumpulan dokumen yang dianalisis.

$$C = S \times M \times P \times \Sigma$$

Dimana:

$C$  = *Coherence Score*  
 $S$  = *Segmentation*  
 $M$  = *Confirmation Measure*  
 $P$  = *Probability Estimation*  
 $\Sigma$  = *Aggregation*

Rumus untuk mencari nilai  $S$  atau *Segmentation* di bawah ini

$$S_{set}^{one} = \{(w^1, w^*) | w^1 = \{w_i\}; w_i \in w; w^* = w\} \dots (2)$$

Rumus untuk mencari nilai  $M$  atau *Confirmation Measure* di bawah ini.

$$NPMI(w', w^*) = \frac{\log \frac{P(w', w^*) + \epsilon}{P(w') P(w^*)}}{-\log(P(w', w^*) + \epsilon)} \dots (3)$$

Rumus untuk mencari nilai  $P$  atau *Probability Estimation* di bawah ini.

$$P_{sw}(w_i, w_j) = \frac{\text{Number of documents in which } w_i \text{ and } w_j \text{ co-occur}}{\text{Number of documents in which } w_i \text{ appears}} \dots (4)$$

Rumus untuk mencari nilai  $\Sigma$  atau *Aggregation* di bawah ini.

$$\Sigma = \frac{\text{jumlah dari confirmation measure}}{\text{banyak confirmation measure}}$$

## 2.6 Evaluasi Model

Evaluasi model adalah proses untuk menilai kinerja dan efektivitas model yang telah dibangun dalam sebuah sistem, terutama dalam konteks pembelajaran mesin dan analisis data. Evaluasi ini bertujuan untuk memahami sejauh mana model mampu membuat prediksi yang akurat atau menghasilkan hasil yang diharapkan berdasarkan data yang digunakan.

## 3. HASIL PEMBAHASAN

Proses pemodelan LDA dimulai dengan tahap preprocessing untuk mempersiapkan data teks agar siap diproses. Setelah data siap, LDA diterapkan untuk memodelkan topik dengan menentukan distribusi kata dalam setiap topik dan distribusi topik dalam setiap dokumen. Proses ini melibatkan beberapa iterasi untuk menghasilkan model yang stabil.

Evaluasi model LDA dilakukan untuk menentukan seberapa baik topik yang dihasilkan mewakili struktur tersembunyi dalam data. Salah satu metrik yang sering digunakan untuk mengevaluasi kualitas pemodelan topik adalah nilai *coherence*. *Coherence* mengukur tingkat kesamaan atau keterkaitan antara kata-kata dalam topik yang dihasilkan. Nilai *coherence* yang lebih tinggi menunjukkan bahwa kata-kata dalam topik tersebut lebih sering muncul bersama dalam dokumen-

dokumen dalam corpus, sehingga topik tersebut lebih koheren dan relevan.

### 3.1 Hasil Preprocessing

Persamaan matematika dinomori dengan Angka Pada proses preprocessing data didalamnya terdapat beberapa fungsi yang digunakan yaitu *stopword removal*, *stemming*, *case folding*. Sehingga data yang sudah di *preprocessing* dapat memudahkan sistem algoritma *lattern dirichlet allocation* ini berjalan. Berikut adalah contoh data yang sudah di *preprocessing*.

NO	FILE TEXT
1	kudim-pembab kubar tingkat jalan terni supla pangan lin
2	getiraco proyek atal atal alas ftkf lin
3	lin proyek non janggang ranta muidar kapa cikin terni tek
4	pengiro sular tingkat fasilitas labuh dukung lin
5	demangpang? nama cari mukan ya berla infomasi labas media pangeruh lin kessali belus undar proses bangsan lin langsung otap semua proyek
6	omondomasa lin lin maku bosan
7	gaj pampin otortia lin bayar
8	baju adal suku mana pakat event negra warga suku unun konon tanah adal serobot untuk lin tipu tampl
9	opanspende pug helat letus wakli belta proyek lin undur wacua
10	jakawi helter bangat buzannya sekarang yang protes buzannya sudah enggaj dapat gaj kayak otortia lin yang enggaj gaj 11

Gambar 2. Data Bersih

### 3.2 Hasil Evaluasi Nilai Coherence

Dalam proses evaluasi, nilai coherence dihitung untuk setiap iterasi topik yang dilakukan. Hasil evaluasi ini digunakan untuk menentukan jumlah topik yang optimal dengan memilih iterasi yang memiliki nilai coherence tertinggi. Pemilihan jumlah topik yang optimal ini sangat penting karena mempengaruhi kualitas akhir dari pemodelan topik. Model LDA yang dihasilkan dengan nilai coherence tertinggi diharapkan mampu memberikan representasi topik yang lebih akurat dan informatif.

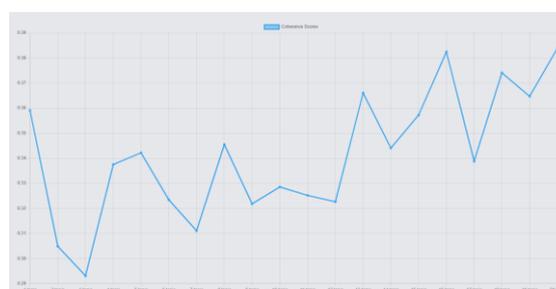
Secara keseluruhan, metode LDA dengan evaluasi nilai coherence memberikan kerangka kerja yang efektif untuk mengidentifikasi dan memahami topik-topik utama dalam kumpulan data teks yang besar. Evaluasi menggunakan nilai coherence memastikan bahwa topik-topik yang dihasilkan tidak hanya matematis signifikan tetapi juga relevan dan koheren dalam konteks data yang dianalisis

Tabel 1. Nilai Coherence

Num Topic	Coherence Score
1	0,35
2	0,30
3	0,29
4	0,33
5	0,34
6	0,32
7	0,31
8	0,34
9	0,32
10	0,32
11	0,32
12	0,32
13	0,36
14	0,34
15	0,35
16	0,38
17	0,33
18	0,37
19	0,36
20	0,38

Dari Tabel 1, terlihat bahwa iterasi topik ke-20 menghasilkan nilai coherence tertinggi. Hal ini menunjukkan bahwa dua puluh topik merupakan jumlah yang paling optimal untuk memberikan interpretasi atau pemahaman yang baik terhadap topik yang dihasilkan. Oleh karena itu, iterasi dengan dua puluh topik akan digunakan kembali dalam proses pemodelan untuk mendapatkan distribusi topik dalam dokumen serta distribusi kata dalam topik berdasarkan jumlah topik tersebut.

Dapat di gambarkan pada gambaran umum, untuk memperjelas hasil dari pemodan topic dengan menggunakan metode algoritma *lattern dirichlet allocation* dan pengujian *coherence score*, dan dipatkan nilai coherence tertinggi pada iterasi topik ke-20 dengan nilai 0.38389228712614887. Seperti yang dapat dilihat pada gambar 3.



Gambar 3. Grafik Coherence Score

### 3.3 Hasil Distribusi dengan LDA Gibbs Sampling

Pemodelan topik dengan menggunakan LDA dan Gibbs Sampling menghasilkan 217 kelompok distribusi topik-dokumen serta 20 kelompok distribusi kata-topik. Setiap kelompok mengandung istilah-istilah dengan bobot yang dijelaskan pada Tabel 2 dan Tabel 3. Bobot ini merepresentasikan probabilitas kemunculan istilah tersebut dalam setiap kelompok. Pada distribusi topik-dokumen, terdapat sampel dari seluruh dokumen yang mencakup kolom 'Topik', yang memuat istilah-istilah dengan bobot dari masing-masing kelompok dokumen. Sementara itu, dalam distribusi kata-topik, terlihat seluruh jumlah topik hasil evaluasi model dengan kolom 'Kata', yang memuat istilah-istilah dengan bobot dari masing-masing kelompok topik. Dapat di lihat seperti contoh tabel di bawah ini.

Tabel 2. Distribusi Topik-Dokumen

Dokumen	Topik	Bobot
Dokumen 1	Topik_16	1
Dokumen 2	Topik_7	0,6
	Topik_5	0,2
	Topik_12	0,2
Dokumen 4	Topik_1	0,5
	Topik_14	0,5

Pada tabel di bawah ini menjelaskan data terkait ditribusi kata-topik yang dimana terdapat nilai bobot pada seluruh kata yang ada di setiap topik. Seperti yang dapat kita lihat contohnya pada tabel 3.

Tabel 3. Distribusi Kata-Topik

Topik	Kata	Bobot
Topik 1	Ikn	0,43
	Proyek	0,34
	Bangun	0,09
	Bayar	0,04
	Paksa	0,04
Topik 3	Rakyat	0,02
	Rakyat	0,33
	Ikn	0,27
	Yang	0,16
	Eggak	0,11
Topik 4	Sudah	0,08
	Air	0,02
	Jabat	0,41
	Mundur	0,29
	Ya	0,20
Topik 9	Banget	0,04
	Eggak	0,04
	Jokowi	0,02
	Saja	0,29
	Ya	0,04
Topik 20	Bangun	0,04
	Jabat	0,04
	Kalo	0,04
	Ikn	0,40
	Bangun	0,23
	Rakyat	0,14
	Nusantara	0,08
	Cnnindonesia	0,06
	Yang	0,06

### 3.4 Hasil Pelabelan

Dalam proses evaluasi, nilai coherence dihitung untuk setiap iterasi topik yang dilakukan. Hasil evaluasi ini digunakan untuk menentukan jumlah topik yang optimal dengan memilih iterasi yang memiliki nilai coherence tertinggi. Pemilihan jumlah topik yang optimal ini sangat penting karena mempengaruhi kualitas akhir dari pemodelan topik. Model LDA yang dihasilkan dengan nilai coherence tertinggi diharapkan mampu memberikan representasi topik yang lebih akurat dan informatif. Berikut adalah contoh beberapa topik yang membahas tentang IKN di program yang sudah dibuat.

## 4. KESIMPULAN

Berdasarkan hasil uji coba dan analisis yang telah dilakukan, penerapan pemodelan topik dengan metode Latent Dirichlet Allocation (LDA) pada 216 data tweet tentang IKN menunjukkan hasil yang positif. Penelitian ini berhasil memberikan gambaran umum mengenai pemodelan topik terkait IKN, di mana pengujian menggunakan nilai coherence menghasilkan nilai tertinggi sebesar 0.38389228712614887 pada topik ke-20 dari 20 topik. Untuk meningkatkan efektivitas aplikasi ini, beberapa saran yang dapat dipertimbangkan meliputi penggabungan algoritma LDA dengan algoritma lain untuk mencapai hasil yang lebih optimal, penerapan confusion matrix untuk menghitung nilai accuracy dalam pemodelan topik, serta pengujian terlebih dahulu dalam pembobotan kata untuk menentukan apakah bag-of-words atau tf-idf yang lebih baik

digunakan. Kesimpulan ini sejalan dengan tujuan dan rumusan masalah yang diangkat dalam penelitian.

Tabel 4. Hasil Pelabelan Distribusi Kata-Topik

Topik	Kata	Jumlah
Proyek Infrastruktur	Ikn	11
	Proyek	
	Bangun	
	Bayar	
	Paksa	
Kebijakan Pemerintah	Kepala	13
	Otorita	
	Mundur	
	Ikn	
	Bambang	
Pendanaan Anggaran	Dana	6
	Eggak	
	Bangun	
	Sudah	
	Perintah	
Ekonomi Investasi	Ikn	18
	Bangun	
	Rakyat	
	Nusantara	
	Cnnindonesia	
Pendanaan Anggaran dan Perumahan Pemukiman	Ikn	6
	Indonesia	
	Otorita	
	Bayar	
	Tapera	

## DAFTAR PUSTAKA

- [1] A. M. Dawis *et al.*, "Digital Signage As Interactive Information Of Nurul Iman Mosque Based On Responsive Web," vol. 1, pp. 8–16, 2023.
- [2] F. Nisrina, "Implementasi Deteksi Topik Putusan Hakim Dengan Latent Dirichlet Allocation (Lda)," pp. 1–64, 2020, [Online]. Available: <https://dspace.uui.ac.id/handle/123456789/23847>
- [3] A. Anas Azhar, "Analisis Pemanfaatan Aplikasi Gojek Berbasis Sistem Informasi Geografis (Persepsi Pelayanan Gojek Pada Mahasiswa Uinsu)," *J. Geogr.*, vol. 12, no. 02, p. 96, 2020, doi: 10.24114/jg.v12i02.17214.
- [4] Y. Afandi, "Gereja dan Pengaruh Teknologi Informasi 'Digital Ecclesiology,'" *Fidei J. Teol. Sist. dan Prakt.*, vol. 1, no. 2, pp. 270–283, 2019, doi: 10.34081/270033.
- [5] 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.
- [6] Ripaldi, E. M. Lantigimo, I. Sasirais, and G. T. Yohannes, "Kesiapan Dayak Deah Di Desa Kinarum Dalam Menyambut Ibu Kota Negara Baru," *J. Teol. Pabelum*, vol. 2, no. 2, pp. 34–48, 2022, doi: 10.59002/jtp.v2i2.27.
- [7] F. N. Hikmah, S. Basuki, and Y. Azhar, "Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation (LDA)," *J. Repos.*, vol. 2, no. 4, pp. 415–426, 2020, doi: 10.22219/repositor.v2i4.52.
- [8] E. Laksono, A. Basuki, and F. Bachtiar, "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no.

- 2, pp. 377–383, 2020, doi: 10.29207/resti.v4i2.1845.
- [9] T. A. Sari, E. Sinduningrum, and F. Noor Hasan, “KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Ulasan Pelanggan Pada Aplikasi Fore Coffee Menggunakan Metode Naïve Bayes,” *Media Online*, vol. 3, no. 6, pp. 773–779, 2023, doi: 10.30865/klik.v3i6.884.
- [10] N. Nurzaman, N. Suarna, and W. Prihartono, “Analisis Sentimen Ulasan Aplikasi Threads Di Google Playstore Menggunakan Algoritma Naïve Bayes,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 967–974, 2024, doi: 10.36040/jati.v8i1.8708.
- [11] N. H. Hrp, M. Fikry, and Y. Yusra, “Algoritma Stemming Teks Bahasa Batak Angkola Berbasis Aturan Tata Bahasa,” *J. Comput. Syst. Informatics*, vol. 4, no. 3, pp. 642–648, 2023, doi: 10.47065/josyc.v4i3.3458.
- [12] M. U. Albab, Y. Karuniawati, and M. N. Fawaiq, “Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic,” *J. Transform.*, vol. 20, no. 2, pp. 1–10, 2023, [Online]. Available: <https://journals.usm.ac.id/index.php/transformatika/page1>
- [13] S. Roiqoh, B. Zaman, and K. Kartono, “Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Mobile JKN dengan Lexicon Based dan Naïve Bayes,” *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1582–1592, 2023, doi: 10.30865/mib.v7i3.6194.