

# Topic Modeling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Menggunakan Metode Latent Dirichlet Allocation

Nathanael Novarian<sup>1\*</sup>, Siti Khomsah<sup>2</sup>, Amalia Beladinna Arifa<sup>3</sup>

<sup>1,3</sup>*Informatic Department, Institut Teknologi Telkom Purwokerto, Indonesia*  
<sup>2</sup>*Data Science Department, Institut Teknologi Telkom Purwokerto, Indonesia*

\*Corresponding Author: [17102112@ittelkom-pwt.ac.id](mailto:17102112@ittelkom-pwt.ac.id)

## Abstract

Telkom Purwokerto (ITTP), students who are enrolled in the Faculty of Information Technology often utilize the university repository to access references while conducting their research or explore research topics. The repository serves as a facility that archives all thesis documents produced by ITTP students. The problem is that the repository does not provide any information or summary regarding the research topic trends that are being wrote. Consequently, students are required to read through each thesis report to identify previous research topics. This research aims to identify the current trends in research topics pursued by ITTP's Faculty of Information Technology students. To achieve our aim, Latent Dirichlet Allocation (LDA) was utilized to build a topic model. The dataset was obtained by crawling abstract documents of theses that are archived in the repository. The coherence value was used to determine the number of topics. Based on the LDA topic modeling analysis of abstracts of theses produced by Faculty of Information Technology students in 2022, 8 topics were identified with a coherence value of 0.446752. Analysis of the top three terms in each topic revealed that the final projects of ITTP's Faculty of Information Technology students in 2022 were related to various themes such as user experience (satisfaction), e-commerce, Covid-19, water, disease detection, case studies, videos, games, learning, networking, funds, tourism, products, Banyumas, villages, scaling techniques, Purbalingga, surveys, environment, homes, and youth.

**Keywords:** *topic modeling; latent dirichlet allocation; tf-idf; repository;thesis*

## Abstrak

Laporan tugas akhir adalah karya hasil penelitian ilmiah mahasiswa strata satu mahasiswa Fakultas Informatika IT Telkom Purwokerto Institut Teknologi Telkom Purwokerto (ITTP). Mahasiswa seringkali menggunakan repositori kampus sebagai acuan referensi tugas akhir mereka. Repositori juga dapat digunakan oleh mahasiswa untuk mengetahui topik apa saja yang banyak diangkat sebagai tugas akhir mahasiswa dikampus. Permasalahannya, ringkasan tren topik-topik penelitian mahasiswa yang ada dalam repositori belum ada sehingga mahasiswa perlu membaca satu persatu laporan tugas akhir untuk mengetahui topik-topik yang sudah pernah diteliti sebelumnya. Pada penelitian bertujuan mengetahui tren topik penelitian mahasiswa FIF ITTP. Metode yang digunakan adalah *Latent Dirichlet Allocation* (LDA) untuk membangun model pengelompokan topik. Dataset yang digunakan adalah abstraksi tugas akhir mahasiswa Fakultas Informatika pada repositori perpustakaan IT Telkom Purwokerto. Metode evaluasi jumlah topik menggunakan metode *coherence value*. Hasil Pemodelan topik dengan LDA pada dataset abstraksi Tugas Akhir mahasiswa FIF pada tahun 2022 menghasilkan 8 topik, didukung dengan nilai coherence sebesar 0.446752. Berdasarkan 3 *term* teratas dari masing-masing kelompok dapat disimpulkan bahwa tugas akhir mahasiswa FIF pada 2022 terkait dengan tema *user experience* (kepuasan), *e-commerce*,

siswa, covid, air, deteksi atau studi kasus terkait penyakit, video, game, pembelajaran, jaringan, belajar, dana, wisata, produk, banyumas, desa, skala, teknik, game, purbalingga, survei, lingkungan, rumah, dan remaja.

**Kata Kunci:** *topic modeling; latent dirichlet allocation; tf-idf; repository; tugas-akhir*

## I. PENDAHULUAN

Repository merupakan penyimpanan dokumen digital untuk menyimpan berbagai dokumen. Salah satu pemanfaatan repository di universitas biasanya untuk menyimpan data-data naskah laporan tugas akhir mahasiswa. Repository dapat membantu pihak pustakawan dalam pendataan laporan tugas akhir dan juga memudahkan mahasiswa dalam melakukan pencarian penelitian yang relevan. Mahasiswa dapat mengakses penelitian-penelitian mahasiswa terdahulu dalam repository untuk mencari referensi judul penelitian yang hendak diajukan. Saat mahasiswa mencari topik-topik penelitian yang terdapat di repository guna mencari referensi penelitian, mahasiswa perlu membaca satu persatu publikasi ilmiah dari seluruh author (penulis). Cara tersebut tentu akan memerlukan usaha, waktu lama dan tenaga. Oleh karena itu, diperlukan suatu metode yang dapat mengotomisasi tren topik penelitian di repository Institut Teknologi Telkom Purwokerto (ITTP).

Langkah otomatisasi yang dapat dilakukan untuk mencari sebuah topik penelitian menggunakan pendekatan *topic modelling*. Beberapa metode untuk menentukan *topic modelling*, salah satunya Latent Dirichlet Allocation (LDA)[1]. LDA adalah model probabilistik generatif dan algoritma unsupervised learning dimana dokumen direpresentasikan sebagai campuran acak topik-topik. Topik- topik tertentu dicirikan oleh distribusi kata-kata dari dokumen-dokumen yang menjadi korpus. Prinsip LDA mengklasifikasikan dokumen sesuai topik terkait. Prinsip algoritma LDA bahwa setiap kata memiliki makna semantik. Sedangkan metode topik modeling lainnya yaitu Latent Semantic Analysis (LSA) yang juga menggunakan prinsip analisis semantik[2]. Berbeda dengan LDA, LSA mengklasifikasi dokumen menurut topik terkait. Prinsip LSA yaitu menghitung jarak vektor untuk mengklasifikasi dokumen sesuai topik terkait. Penggunaan LSA juga jarang ditemui dibanding penggunaan LDA. Beberapa penelitian menerapkan LDA untuk mencari topik dalam kumpulan korpus berbahasa Indonesia [1]–[10]. Penelitian [3]menerapkan LDA untuk menemukan topik-topik penelitian dari korpus keperawatan yang diperoleh dari OJS jurnal SINTA. Sedangkan penelitian [8] juga menerapkan LDA untuk menemukan topik-topik penelitian para dosen di jurnal JPTEI UNY. Hasil penerapan LDA pada penelitian[8] dan [3] mampu mengenali jenis-jenis topik dari kumpulan korpus dokumen hasil penelitian.

Berdasarkan latar belakang diatas maka penelitian bertujuan mengetahui topik-topik penelitian mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto. Metode untuk pemodelan yang digunakan adalah LDA , sedangkan vektorisasi dokumen menggunakan *term-frequency invers document frequency* (TF-IDF).

## II. LITERATURE REVIEW

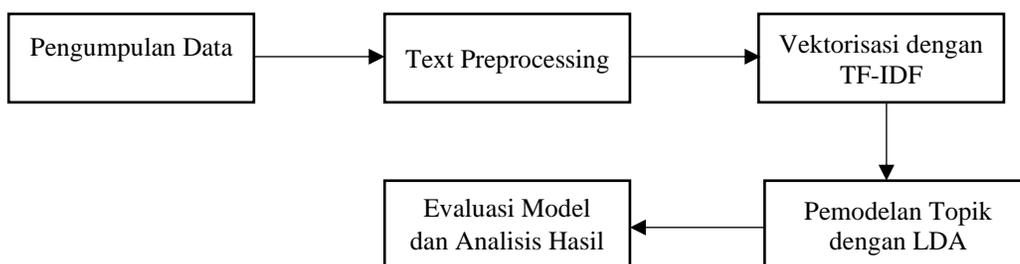
*Topic modeling* bertujuan mengesktrak hubungan antar satu kata dengan kata lainnya dalam kumpulan dokumen. Metode *topic modeling* bisanya digunakan untuk memahami topik- topik pembicaran yang populer yang dapat diungkap berdasarkan korpus tertentu misalnya bidang kesehatan, pendidikan, peneitian, dan lain sebagainya. Penelitian *topic modeling*, korpus bahasa sudah banyak dilakukan oleh para peneliti. Beberapa penerapan diantaranya, penelitian [8]menerapkan *latent dirichlet allocation* (LDA) untuk mencari tren topik penelitian dosen pada jurnal JPTEI UNY. Korpus yang digunakan diambil dari Google Scholar. Model yang dibangun dapat menghasilkan informasi bahwa publikasi yang terdapat pada jurnal JPTEI UNY adalah seputar topik pendidikan vokasi, pengembangan sistem, media pembelajaran, dan sistem pembelajaran di SMK. Penelitian [3]menggunakan korpus dokumen publikasi ilmiah berupa Tugas Akhir mahasiswa keperawatan. Penelitian tersebut juga menggunakan LDA. Penelitian ini bertujuan untuk menganalisis tren topik penelitian dengan pemodelan topik terhadap judul-judul penelitian di bidang keperawatan pada repository jurnal SINTA. Representasi dokumen yang digunakan yaitu TF-IDF. Model yang dibangun menghasilkan topik tentang ibu hamil, fenomenologi, tingkat kecemasan, penderita hipertensi, tingkat stress, intensitas nyeri, penderita

diabetes, kesehatan reproduksi, ibu nifas, kualitas tidur, asuhan keperawatan, tekanan darah, kinerja perawat, kualitas hidup, relaksasi otot dan lain-lain. Penelitian [9] menggunakan korpus abstraksi skripsi mahasiswa yang berasal dari 584 judul. Penelitian ini bertujuan untuk menganalisis tren pada topik skripsi mahasiswa program studi sastra inggris di Universitas Islam Negeri Sunan Ampel Surabaya (UINSA). Representasi dokumen yang digunakan yaitu *Bag of Words* atau jumlah kemunculan kata. Penentuan jumlah topik menggunakan metode percobaan, mulai dari dua sampai lima topik. Setiap percobaan menggunakan iterasi 100, 500, 1000, dan 5000 untuk mendapatkan model yang fit. Kesimpulannya, model yang fit yaitu jika topiknya adalah tiga. Setelah hasil tren topik diverifikasi oleh stakeholder (Program Studi Sastra Inggris UINSA) ternyata ada satu topik yang menghasilkan tren yang sesuai Program Studi Sastra Inggris UINSA. Penelitian menggunakan korpus sekumpulan laporan yang telah diberi label. Penelitian ini bertujuan untuk mengetahui topik tersembunyi pada dokumen pada laporan Tugas Akhir mahasiswa dengan mengimplementasi *gibbs sampling* dan mengetahui relevansi klasifikasi topik yang dihasilkan. Representasi dokumen yang digunakan yaitu TF-IDF. LDA yang digunakan sudah dimodifikasi dengan metode *collapsed gibbs sampling*. Penelitian ini menyimpulkan bahwa metode probabilitas sebuah kata dipengaruhi oleh banyaknya jumlah topik dan dokumen. Penelitian [11] juga menyimpulkan bahwa LDA sensitif terhadap komposisi kata yang mengandung banyak kata umum sehingga mengurangi akurasi. Penelitian [7] menggunakan korpus judul dan abstrak yang dibobotkan dengan TF-IDF dan dikelompokkan menggunakan K-Means Clustering. Penelitian [7] tersebut juga menggunakan LDA-*gibbs sampling* dengan tujuan membuat sistem rekomendasi topik tugas akhir berdasarkan kompetensi dalam transkrip akademik. Keimpulannya, semakin tinggi nilai kemiripan vektor kata inti dari sistem akan memberikan nilai probabilitas yang tinggi dalam topik terpilih [7].

Penelitian- penelitian tersebut diatas menggunakan metode LDA untuk membuat model topik. Sedangkan TF-IDF seringkali digunakan sebagai ekstraksi fitur. Sedangkan analisis jumlah kelompok topik yang tepat umumnya menggunakan percobaan dengan menentukan jumlah topik tertentu. Penentuan jumlah topik yang tepat secara manual dapat menghasilkan kesimpulan yang bias. Oleh karena itu, penelitian ini akan menggunakan metode *coherence value* untuk menemukan jumlah topik yang tepat.

### III. RESEARCH METHODS

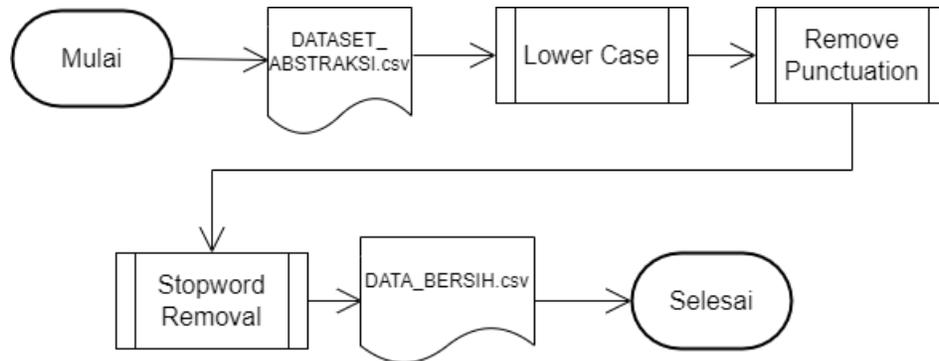
Subyek penelitian ini merupakan data dari abstraksi tugas akhir mahasiswa fakultas informatika pada tahun 2022 yang diambil dari repositori Institut Teknologi Telkom Purwokerto (ITTP). Sedangkan obyek penelitian ini adalah topik modeling dari abstraksi tugas akhir mahasiswa di fakultas informatika pada tahun 2022. Penelitian ini dilakukan sesuai dengan tahapan pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Pengumpulan data.  
Data diambil secara manual dari *website* repositori ITTP yang berjumlah 428 abstraksi. Abstraksi pada Tugas Akhir merupakan rangkuman dari inti topik pembicaraan pada penelitian Tugas Akhir tersebut.
2. Text Preprocessing

Tahapan *text preprocessing* ditunjukkan pada Gambar 2 yaitu terdiri proses load data abstraksi dalam format csv, mengubah menjadi huruf kecil, menghapus tanda baca, menghilangkan *stopword*, selanjutnya hasil data akan disimpan dalam file *data\_bersih.csv*.



Gambar 1. Text Preprocessing

Gambar 2 adalah tahapan *text preprocessing* terdiri dari proses *lower case*, *remove punctuation*, dan *stopword removal*. *Text preprocessing* bertujuan membersihkan data sehingga bebas dari. *Lower case* adalah proses untuk merubah huruf kapital yang terdapat di suatu kata/kalimat menjadi huruf kecil. *Remove punctuation* adalah proses untuk membersihkan data dari tanda baca seperti “, - , \_ , ! , % , ? , ( , & , ^ , \$ , # , @ , + , > , < ” . Proses ini juga akan menghapus angka “0-9”. *Stopword removal* adalah proses untuk menghilangkan kata-kata yang tidak memiliki makna. Terdapat beberapa kata yang tidak memiliki makna seperti kata hubung “yang”, “dari”, “ke”, “di”, “pada”, “dimanapun”, dan lain-lain. *Stopword removal* menggunakan *library stopwords NLTK* yang berisi 765 kata *stopword*.

3. Vektorisasi (TF-IDF)

Sebelum membangun model LDA, diperlukan adanya tahap vektorisasi. Vektorisasi atau pembobotan kata pada penelitian ini menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF dihitung dengan menggunakan persamaan (1).

$$W_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right) \dots \dots \dots (1)$$

Dimana ,  $W_{i,j}$  adalah bobot TF-IDF term / kata ke-i dalam dokumen ke-j. Sedangkan  $tf_{i,j}$  adalah frekuensi kata ke-i pada dokumen j. Sedangkan N adalah total dokumen dalam korpus yang digunakan dan  $df_i$  adalah jumlah dokumen yang mengandung *term* ke-i

4. Topic Modeling Menggunakan LDA

Cara kerja LDA adalah menghitung *joint probability distribution* dengan cara melakukan sampling satu persatu terhadap setiap variabel lainnya alias *full conditional probability*. Dalam pengelompokkan topik memiliki dua bentuk distribusi probabilitas yang harus dicari yaitu, distribusi probabilitas dokumen pada suatu dokumen dan distribusi probabilitas kata pada suatu topik[10]. Rumus LDA menggunakan persamaan (2).

$$\rho(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \dots \dots \dots (2)$$

Dimana :  
 M adalah jumlah dokumen;  
 N adalah jumlah kata dalam dokumen tertentu;  
 $\alpha$  adalah *dirichlet prior* pada sebaran topik per dokumen;

$\beta$  adalah *dirichlet prior* pada sebaran kata per topik;  
 $\theta_i$  adalah sebaran topik untuk dokumen  $i$ ;  
 $\varphi_k$  adalah sebaran kata untuk topik  $k$ ;  
 $z_{ij}$  adalah topik untuk kata ke- $j$  pada dokumen  $i$ ;  
 $w_{ij}$  adalah kata spesifik.

#### 5. Evaluasi Model dan Analisis Hasil.

Analisis topik dilakukan dengan menganalisis topik-topik tiap kelompok dalam model LDA yang dihasilkan. Untuk mendapatkan jumlah kelompok, metode *coherence value* digunakan untuk menentukan jumlah kelompok yang optimal. Setelah dapat jumlah kelompok didapatkan, kemudian model LDA akan dibangun dengan parameter jumlah topik sesuai hasil *coherence value*. Analisis *term* (kata) tiap kelompok topik menggunakan PyLDAvis sebagai alat visualisasi.

### IV. RESULTS AND DISCUSSION

Penelitian ini mengambil data secara manual dari *website* <https://repository.itelkom-pwt.ac.id/>. Data yang diambil adalah data abstraksi dari tugas akhir mahasiswa informatika pada tahun 2022. Jumlah data yang diperoleh dari hasil pengumpulan data ini sebanyak 428 abstraksi. Dataset kemudian dijadikan korpus untuk membuat *topic modeling*. Dataset dibersihkan terlebih dahulu kemudian diubah menjadi vektor TF-IDF. Vektor TF-IDF kemudian menjadi masukan model LDA. Pada pemodelan dengan LDA, diperlukan masukan parameter yaitu jumlah topik. Dalam penelitian ini, jumlah topik ditentukan mencari nilai *coherence value*-nya.

#### 1. Nilai *Coherence*

Nilai *coherence* digunakan untuk mengevaluasi model topik. Model yang baik akan mempunyai nilai *coherence* yang tinggi. Komputasi nilai *coherence*, didapatkan dengan fungsi *compute\_coherence* berikut:

```
def compute_coherence_values(dictionary, corpus, texts, limit, start, step):
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit, step):
        model = LdaModel(corpus=corpus,
                        id2word=dictionary, num_topics=num_topics, iterations=100)
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model,
                                       texts=texts, dictionary=dictionary, coherence='c_v')
        coherence_values.append(coherencemodel.get_coherence())
    return model_list, coherence_values
start=1, limit=20, step=1
model_list, coherence_values = compute_coherence_values(dictionary, corpus=corpus_tfidf, texts=hasil_stop,
start=start, limit=limit, step=step)
```

Kemudian, hasil komputasi *coherence* divisualkan menggunakan kode program berikut:

```
#show graphs
import matplotlib.pyplot as plt
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

Nilai *coherence* perlu diujicoba berkali-kali dan setiap hasilnya akan dibandingkan, kemudian akan dilihat nilai yang paling tinggi. Pada penelitian ini menggunakan 20 kali percobaan sehingga ada 20 kali komputasi nilai *coherence*. Jumlah percobaan ini mengindikasikan jumlah topik yang akan dianggap terbaik

bagi model yang dibangun. Nilai maksimal tersebut digunakan sebagai nilai rujukan untuk menentukan jumlah topik yang dianggap paling baik. Detil nilai *coherence* setiap topik ditunjukkan Gambar 3, dan perbandingan dengan Grafik ditunjukkan Gambar 4.

```

Num Topics = 1 has Coherence Value of 0.363506
Num Topics = 2 has Coherence Value of 0.358853
Num Topics = 3 has Coherence Value of 0.334181
Num Topics = 4 has Coherence Value of 0.361393
Num Topics = 5 has Coherence Value of 0.399079
Num Topics = 6 has Coherence Value of 0.376111
Num Topics = 7 has Coherence Value of 0.39255
Num Topics = 8 has Coherence Value of 0.446752
Num Topics = 9 has Coherence Value of 0.434875
Num Topics = 10 has Coherence Value of 0.414407
    
```

Gambar 3. Daftar Nilai Coherence 20 Percobaan (20 topik)

Berdasarkan Gambar 3, *coherence value* tertinggi ada pada iterasi topik ke-8 dengan nilai 0.446752. Berdasarkan hasil tertinggi nilai *coherence* tersebut, jumlah topik yang akan digunakan sebagai masukan parameter *num\_topic* pada model LDA adalah 8

## 2. Hasil Ekstraksi Topik dengan LDA

Pemodelan topik menggunakan LDA menghasilkan 8 kelompok topik. Masing-masing kelompok topik mengandung *term* beserta bobotnya, pada Tabel 1. Setiap *term* dalam kelompok topik memiliki bobot yang merepresentasikan probabilitas kemunculan *term* tersebut dalam keseluruhan dokumen dataset.

Tabel 1. Topik, word (term), dan bobotnya

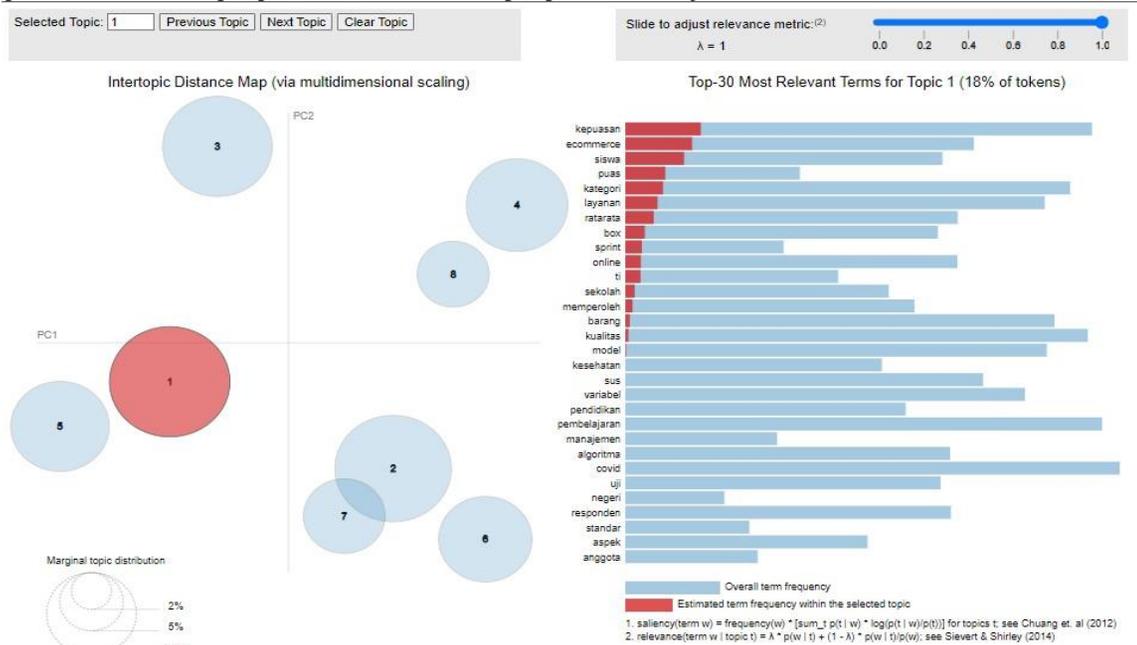
Topik	Word	Bobot
Topik 1	kepuasan	0.005
	ecommerce	0.005
	siswa	0.005
	puas	0.005
	kategori	0.005
	layanan	0.004
	ratarata	0.004
	box	0.004
	sprint	0.004
	online	0.004
Topik 2	air	0.006
	penyakit	0.006
	covid	0.005
	sensor	0.004
	jawa	0.004
	it	0.004
	bayes	0.004
	wisata	0.004
	interface	0.004
	desain	0.004
Topik 3	video	0.006
	pembelajaran	0.005
	game	0.005
	penyakit	0.005
	program	0.005
	learning	0.004
	machine	0.004
	cnn	0.004
	akurasi	0.004

Topik	Word	Bobot
Topik 4	kualitas	0.004
	jaringan	0.009
	belajar	0.005
	dana	0.005
	gangguan	0.004
	fitur	0.004
	anak	0.004
	pt	0.004
	barang	0.004
	keamanan	0.004
	desa	0.004
Topik 5	wisata	0.009
	produk	0.006
	banyumas	0.006
	kabupaten	0.005
	model	0.005
	tanaman	0.005
	video	0.004
	poster	0.004
	wisatawan	0.004
	kesalahan	0.004
	Topik 6	desa
skala		0.005
teknik		0.005
variabel		0.004
akurasi		0.004
web		0.004
algoritma		0.004
pemilihan		0.004
pelayanan		0.004
desain		0.004
Topik 7		game
	purbalingga	0.006
	survei	0.005
	algoritma	0.004
	positif	0.004
	signifikan	0.004
	desa	0.004
	desain	0.004
	pembayaran	0.004
	experience	0.004
Topik 8	lingkungan	0.005
	rumah	0.005
	remaja	0.005
	platform	0.005
	game	0.005
	edukasi	0.004
	kompetensi	0.004
	pelanggan	0.004
	mental	0.004
	simple	0.004

### 3. Visualisasi Topik Modeling Menggunakan PyLDAVis

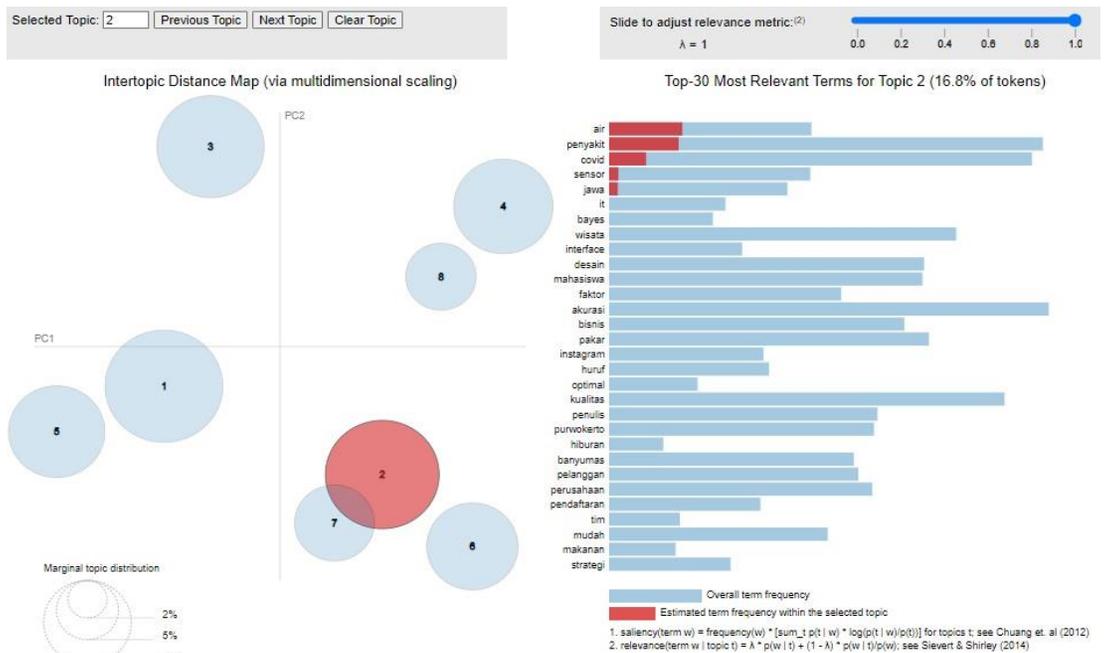
Setelah mendapat bobot kata per topik, maka selanjutnya adalah tahap visualisasi agar dapat dengan mudah dianalisis. Visualisasi menggunakan pyLDAvis dirancang untuk menafsirkan topik dalam model

topik yang sesuai dengan kumpulan data teks. Visualisasi PyLDAvis akan selalu menampilkan 30 term pertama dalam setiap topik. Hasil visualisasi topik pertama ditunjukkan Gambar 4.



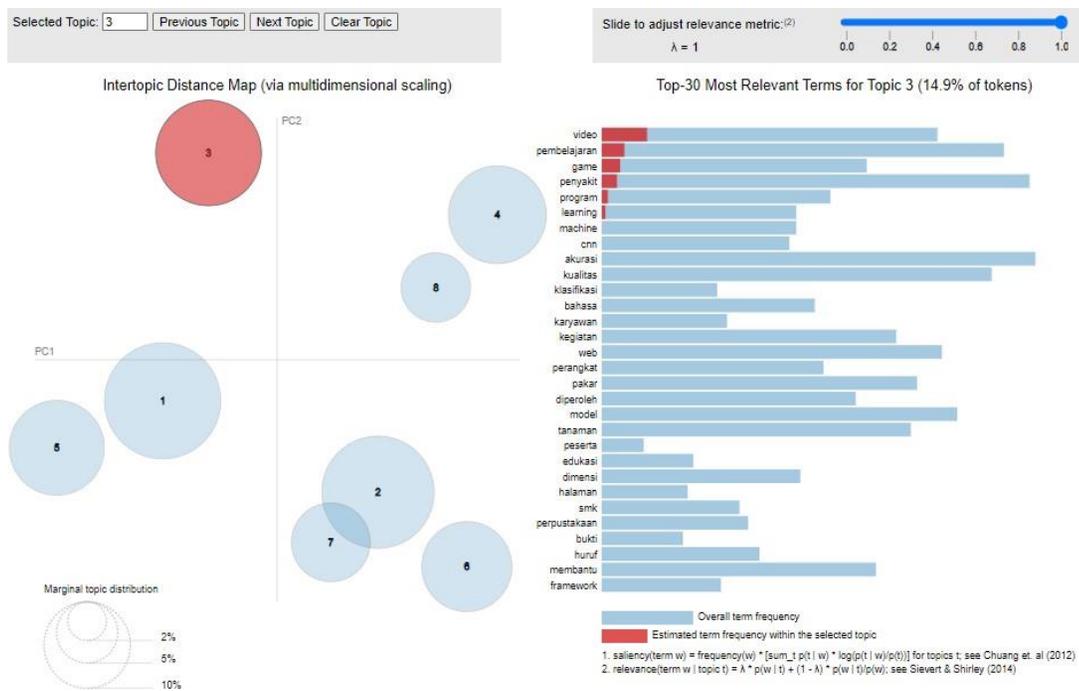
Gambar 4. Topik Pertama

Gambar 4 menampilkan 30 term dari topik pertama namun jika diamati, grafik paling panjang yaitu 10 term teratas yaitu “kepuasan”, "ecommerce", "siswa", "puas", "kategori", "layanan", “ratarata”, "box", "sprint”, "online”. Dari 10 term teratas, kata “kepuasan”, "ecommerce", dan "siswa" menduduki 3 besar kata yang populer pada topik pertama. Ketiga term tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok pertama lebih cenderung terkait dengan kepuasan pengguna yang bisa diasumsikan ranah kajian *user experience*. Kemudian term *e-commerce* dan siswa menunjukkan bahwa tugas akhir terkait *e-commerce* dan siswa.



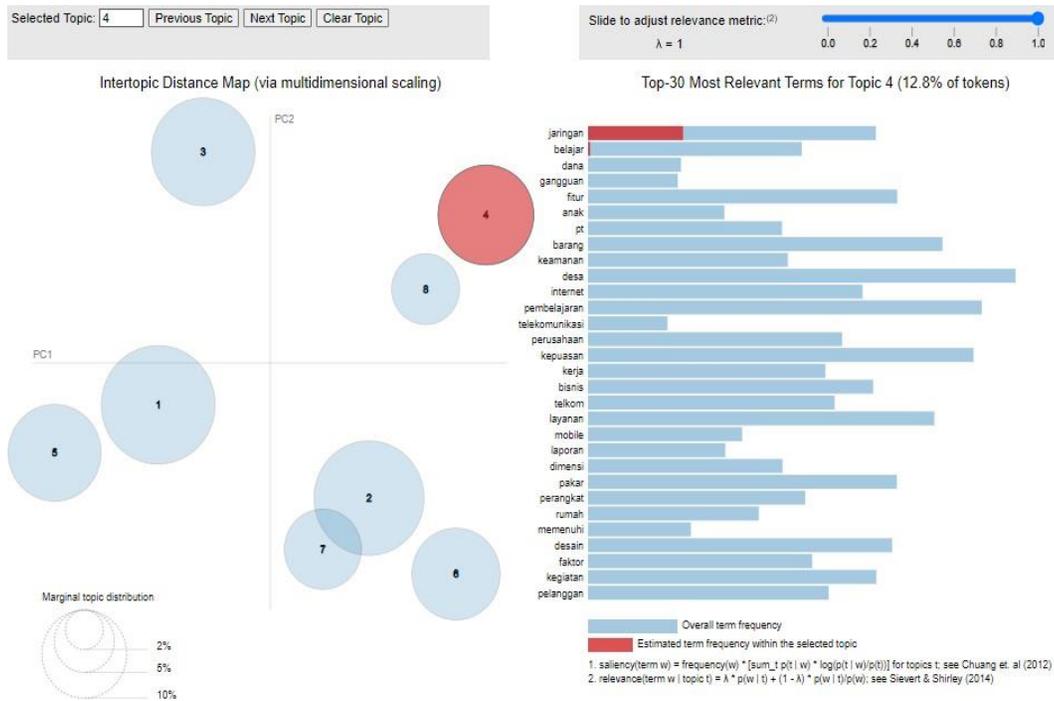
Gambar 5. Topik Kedua

Gambar 5 dapat dilihat bahwa topik kedua memiliki 10 *term* teratas yaitu "air", "penyakit", "covid", "sensor", "jawa", "it", "bayes", "wisata", "interface", "desain". Dari 10 *term* teratas, kata "air", "penyakit", dan "covid" menduduki 3 besar kata yang populer pada topik kedua. Ketiga *term* tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 2 lebih cenderung terkait dengan tema penyakit.



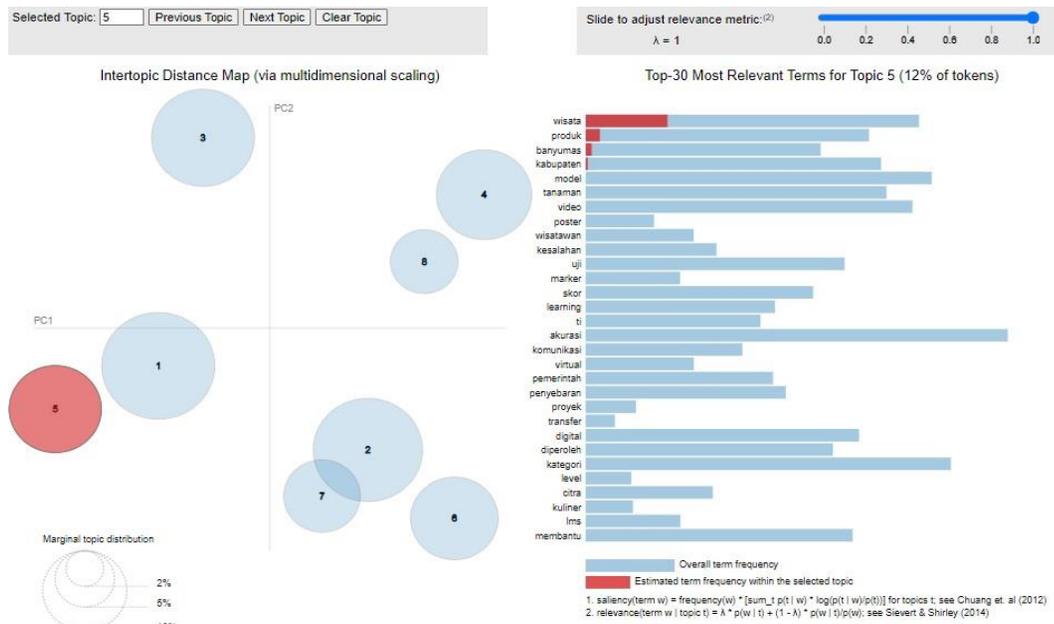
Gambar 6. Topik Ketiga

Gambar 6 dapat dilihat bahwa topik ketiga memiliki 10 term teratas yaitu "video", "pembelajaran", "game", "penyakit", "program", "learning", "machine", "cnn", "akurasi", "kualitas". Dari 10 term teratas, kata "video", "pembelajaran", dan "game" menduduki 3 besar kata yang populer pada topik ketiga. Ketiga term tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 3 lebih cenderung terkait dengan tema video dan game. Term "pembelajaran" bisa jadi menunjukkan bahwa tren topik penelitian mahasiswa tersebut terkait dengan pembelajaran atau edukasi.



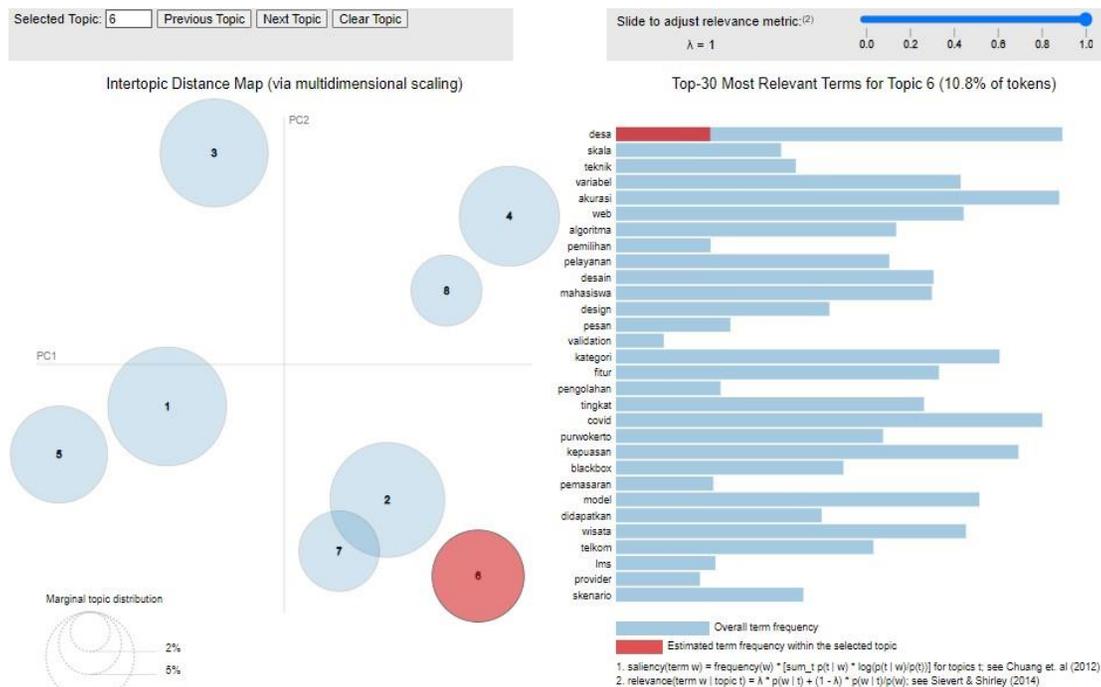
Gambar 7. Topik Keempat

Gambar 7 menunjukkan topik pada kelompok keempat memiliki 10 term teratas yaitu "jaringan", "belajar", "dana", "gangguan", "fitur", "anak", "pt", "barang", "keamanan", "desa". Dari 10 term teratas, kata "jaringan", "belajar", dan "dana" menduduki 3 besar kata yang populer pada topik keempat. Ketiga term tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 4 lebih cenderung terkait dengan tema jaringan. Term "pembelajaran" bisa jadi menunjukkan bahwa tren topik penelitian mahasiswa tersebut terkait dengan edukasi terkait dengan jaringan. Term "dana" diduga menunjukkan tentang bagaimana menghemat pengeluaran anggaran dalam membangun aplikasi.



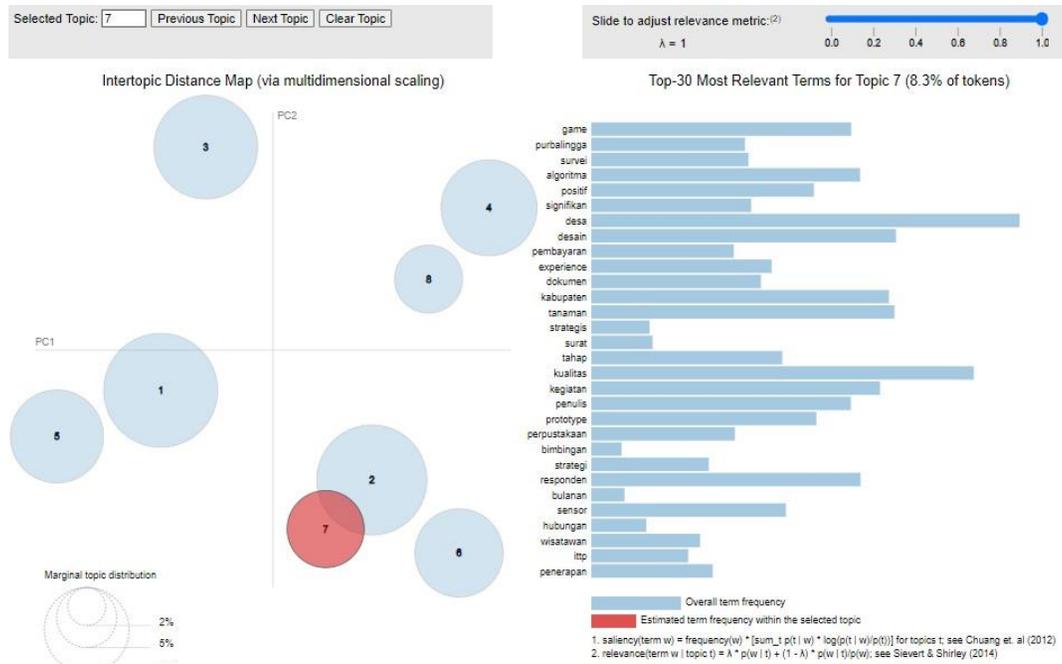
Gambar 8. Topik Kelima

Gambar 8 menunjukkan pada kelompok kelima memiliki 10 topik teratas yaitu “wisata”, “produk”, “banyumas”, “kabupaten”, “model”, “tanaman”, “video”, “poster”, “wisatawan”, “kesalahan”. Dari 10 *term* teratas, kata “wisata”, “produk”, dan “banyumas” 3 kata yang populer pada topik kelima. Ketiga *term* tersebut menunjukkan bahwa topik penelitian mahasiswa pada kelompok 5 lebih cenderung terkait dengan tema wisata. Term “produk” bisa jadi menunjukkan bahwa tren topik penelitian mahasiswa tersebut terkait dengan produk. Term “banyumas” diduga menunjukkan tentang objek penelitian yang ditunjukkan pada kabupaten Banyumas.



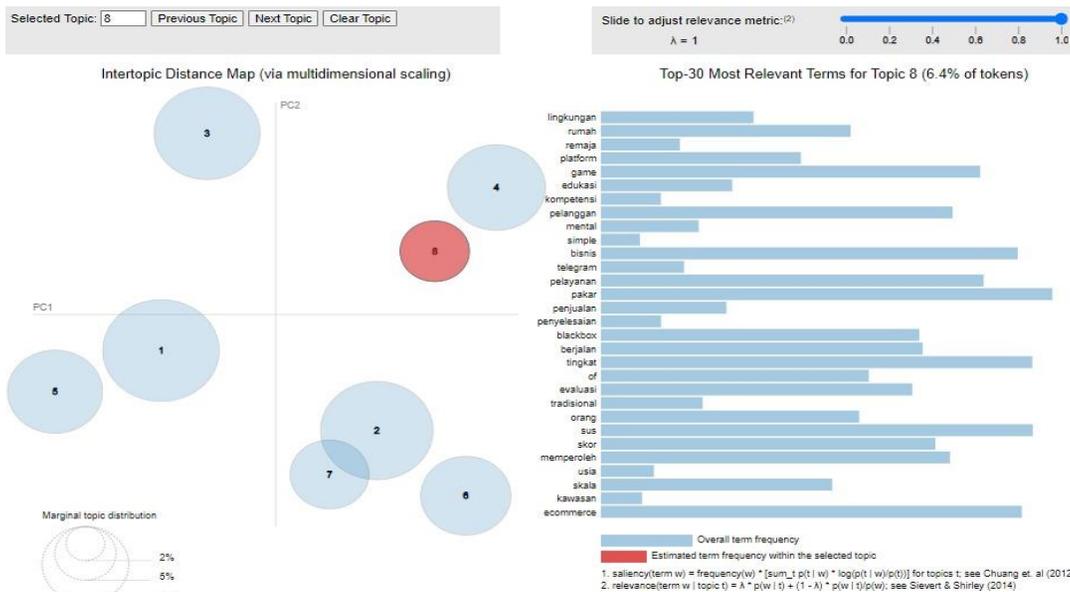
Gambar 9. Topik Keenam

Gambar 9 adalah kelompok topik keenam. Sepuluh (10) *term* teratas yaitu “desa”, “skala”, “teknik”, “variabel”, “akurasi”, “web”, “algoritma”, “pemilihan”, “pelayanan”, “desain”. Sedangkan jika dilihat 3 topik teratas adalah kata “desa”, “skala”, dan “teknik”. Sepuluh *term* tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 6 lebih cenderung terkait dengan web. Term “akurasi” bisa jadi menunjukkan bahwa tren topik penelitian mahasiswa tersebut terkait dengan melihat akurasi hasil penelitian. Term “desa” diduga menunjukkan tentang objek penelitian yang ditunjukkan pada desa.



Gambar 10. Topik Ketujuh

Gambar 10 merupakan kelompok topik ketujuh. Sepuluh teratas yaitu “game”, “purbalingga”, “survei”, “algoritma”, “positif”, “signifikan”, “desa”, “desain”, “pembayaran”, “experience”. Dari 10 term teratas, kata “game”, “purbalingga”, dan “survei” adalah 3 kata yang populer. Sepuluh term tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 7 lebih cenderung terkait dengan *game* dan survei serta subyek atau obyek penelitian Purbalingga.



Gambar 11. Topik Kedepalan

Gambar 11 adalah representasi visual topik kedelapan. Sepuluh *term* teratas yaitu “lingkungan”, “rumah”, “remaja”, “platform”, “game”, “edukasi”, “kompetensi”, “pelanggan”, “mental”, “simple”. Dari 10 *term* teratas, kata “lingkungan”, “rumah”, dan “remaja” adalah 3 kata yang paling sering muncul. Kesepuluh *term* tersebut menunjukkan bahwa topik penelitian mahasiswa itu pada kelompok 8 lebih cenderung terkait dengan remaja, lingkungan, dan rumah.

## V. CONCLUSION

Pemodelan topik dengan LDA pada dataset abstraksi Tugas Akhir mahasiswa FIF pada tahun 2022 menghasilkan 8 topik. Hal tersebut didukung dengan nilai *coherence* sebesar 0.446752. Berdasarkan analisis pada 3 *term* teratas dari setiap kelompok topik tersebut, dapat disimpulkan bahwa tugas akhir mahasiswa FIF pada 2022 terkait dengan tema *user experience* (kepuasan), *e-commerce*, siswa, covid, air, deteksi atau studi kasus terkait penyakit, video, game, pembelajaran, jaringan, belajar, dana, wisata, produk, banyumas, desa, skala, teknik, game, purbalingga, survei, lingkungan, rumah, dan remaja. Tema-tema tersebut ada yang jelas menunjukkan subyek atau obyek namun masih ada *term* yang kurang berkaitan dengan tugas akhir, seperti lingkungan, rumah, dana, belajar, dan air. Penulis menyarankan ada penelitian lebih lanjut untuk tahap pembersihan data sehingga *term* yang tidak umum sebagai tema tugas akhir tidak muncul.

## REFERENCES

- [1] H. Jelodar *et al.*, “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” *Multimed Tools Appl*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.
- [2] A. Verma and A. K. Gahier, “Topic modeling of E-news in Punjabi,” *Indian J Sci Technol*, vol. 8, no. 27, 2015, doi: 10.17485/ijst/2015/v8i27/81665.
- [3] S. Yoga, F. Nurul Isnaini, and O. Pamulatsih Dwi, “Pemodelan Topik Penelitian Bidang Keperawatan Indonesia pada Repository Jurnal Sinta Menggunakan Metode Topic Modelling LDA (Latent Dirichlet Allocation),” *Seminar Nasional Teknologi Informasi & Ilmu Komputer*, vol. 1, pp. 90–102, 2020.
- [4] Y. U. Al-khairi, Y. Wibisono, and B. L. Putro, “Deteksi Topik Fashion pada Twitter dengan Latent Dirichlet Allocation Fashion Topic Detection on Twitter with Latent Dirichlet Allocation,” vol. 1, no. 1, 2018, [Online]. Available: <https://ejournal.upi.edu/index.php/JATIKOM>
- [5] B. W. Arianto and G. Anuraga, “Topic Modeling for Twitter Users Regarding the ‘Ruangguru’ Application,” *Jurnal ILMU DASAR*, vol. 21, no. 2, p. 149, 2020, doi: 10.19184/jid.v21i2.17112.
- [6] C. S. Hudaya, H. Fakhurroja, and A. Alamsyah, “Analisis Persepsi Konsumen Terhadap Brand Go-Jek Pada Media Sosial Twitter Menggunakan Metode Sentiment Analysis Dan Topic Modelling,” *Jurnal Mitra Manajemen*, vol. 3, no. 6, pp. 664–673, 2019, doi: 10.52160/ejmm.v3i6.244.
- [7] D. Purwitasari, Aida Muflichah, Novrindah Alvi Hasanah, and Agus Zainal Arifin, “Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 421–428, 2021, doi: 10.29207/resti.v5i3.3049.
- [8] A. Nurlayli and Moch. A. Nasichuddin, “Topik Modeling Penelitian Dosen Jptei Uny Pada Google Scholar Menggunakan Latent Dirichlet Allocation,” *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 4, no. 2, pp. 154–161, 2019, doi: 10.21831/elinvo.v4i2.28254.
- [9] A. I. Alfanzar, K. Khalid, and I. S. Rozas, “Topic Modelling Skripsi Menggunakan Metode Latent Dirichlet Allocation,” *JSiI (Jurnal Sistem Informasi)*, vol. 7, no. 1, p. 7, 2020, doi: 10.30656/jsii.v7i1.2036.
- [10] A. F. Hidayatullah, M. R. Ma’arif, M. Habibie, and S. Khomsah, “Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation,” *IOP Conf Ser Mater Sci Eng*, vol. 1077, no. 1, p. 012012, 2021, doi: 10.1088/1757-899x/1077/1/012012.
- [11] U. T. Setijohatmo *et al.*, “Prosiding The 11 th Industrial Research Workshop and National Seminar Bandung,” 2020.