JURNAL PINTER

e – ISSN: 2597 – 4475

VOL 8. NO. 2 DESEMBER 2024 http://doi.org/10.21009/pinter.8.1.3

TOPIC MODELING DOKUMEN SKRIPSI PRODI PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER UNIVERSITAS NEGERI JAKARTA MENGGUNAKAN METODE *LATENT DIRICHLET ALLOCATION*

Eki Nugraha[1], Widodo[2], Murien Nugraheni2[3]

[1] Mahasiswa Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ
[2,3] Dosen Prodi Pendidikan Teknik Informatika dan Komputer, Teknik Elektro, FT – UNJ
[1] EkiNugraha_1512617010@mhs.unj.ac.id, [2] widodo@unj.ac.id, [3] muriennugraheni@unj.ac.id

Abstrak

Skripsi atau tugas akhir menjadi suatu syarat untuk memperoleh gelar Sarjana Strata-1 di perguruan tinggi di Indonesia. Pada program studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta mahasiswa mengambil judul dan tema skripsi yang beragam, tidak semua tema skripsi yang dibuat sesuai dengan konsentrasi jurusan masing-masing. Dengan semakin bervariasinya dokumen skripsi maka dilakukan pemodelan topik skripsi dengan menggunakan metode Latent Dirichlet Allocation untuk mengetahui komposisi topik skripsi yang dapat digunakan sebagai referensi pada penelitian selanjutnya. Bahan penelitian yang digunakan berasal dari repository admin Prodi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta berupa softcopy tahun lulus 2017-2022 sebanyak 329 dokumen skripsi. Dokumen skripsi diproses melalui beberapa tahapan yaitu pre-processing yang terdiri dari case folding, stopwords,lemmatization, stemming, setelah itu dilakukan pemodelan LDA dengan menggunakan library Gensim, pembobotan kata dengan TF-IDF, pengujian topik menggunakan coherence score dan perplexity. Sehingga didapatkan 10 topik yang sering dibahas yaitu topik ke-1 mengenai Pengembangan media pembelajaran sebanyak 50 dokumen, topik ke-2 Evaluasi pembelajaran sebanyak 42 dokumen, topik ke-3 Pengembangan System informasi berbasis website sebanyak 42 dokumen, topik ke-4 Teknik komputer dan jaringan sebanyak 36 dokumen, topik ke-5 Pengembangan System informasi Universitas Negeri Jakarta berbasis website menggunakan user interface, user experience ,& backend sebanyak 36 dokumen, topik ke-6 Pengembangan dan evaluasi media pembelajaran e-learning menggunakan algoritma sebanyak 34 dokumen, topik ke-7 data mining sebanyak 28 dokumen, topik ke-8 System requirement pengembangan perangkat lunak sebanyak 26 dokumen, topik ke-9 Standar pencapaian sumber daya Pendidikan sebanyak 14 dokumen, topik ke-10 Evaluasi pengembangan website, multimedia dan jaringan sebanyak 21 dokumen. Pengujian menggunakan coherence score menghasilkan 0.2614 dan perplexity -5.9895.

Kata Kunci: **Pemodelan Topik, LDA, TF-IDF, *Coherence Score, Perplexity***

1. Pendahuluan

Skripsi atau tugas akhir menjadi suatu syarat untuk memperoleh gelar Sarjana Strata-1 di perguruan tinggi di Indonesia, tugas akhir merupakan hasil pengamatan terhadap suatu masalah yang terjadi dengan menggunakan metode tertentu di bidang ilmu tersebut (Mahmud,2019), skripsi biasa disebut untuk mahasiswa yang akan mendapatkan gelar S-1 sedangkan untuk bidang Diploma biasa disebut dengan tugas akhir. Pendidikan Teknik Informatika dan Komputer (PTIK) merupakan salah satu prodi S-1 di Fakultas Teknik Universitas Negeri Jakarta yang berdiri sejak 31 juli 2009 dengan SK Penyelenggaraan Program Studi: 13300/D/T/K-N/2012 dan pertama kali dibuka pada periode 2010/2011. Pada Program Studi ini mempunyai tiga konsentrasi jurusan di antaranya adalah Rekayasa Perangkat Lunak, Teknik Komputer Jaringan, dan Multimedia.

Penulis mencari sumber data dengan membaca dokumen skripsi sebelumnya di *repository* admin Prodi Pendidikan Teknik Informatika Dan Komputer dan *website repository.unj.ac*.id. Pada kenyataannya dalam menyelesaikan studi S-1 mahasiswa Prodi Pendidikan Teknik Informatika dan Komputer tidak semua tema skripsi yang dibuat sesuai dengan konsentrasi jurusan masing-masing, karena ada beberapa mahasiswa yang lebih tertarik di bidang konsentrasi yang berbeda dalam melakukan penelitian karena ketertarikan dengan mata kuliah yang diampu sebelumnya,

Available at:
https://journal.unj.ac.id/unj/index.php/pinter/article/view/53187

Oleh karena itu Prodi Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta belum diketahui jumlah peta topik penelitian skripsi, mengingat semakin lama banyak dokumen skripsi yang masuk, dan masing-masing skripsi yang dibuat mempunyai label sendiri yang tentunya mempunyai topik berbeda-beda dalam satu dokumen, dan teknik pengelompokan hanya dengan melihat topik skripsi secara manual dari judul dan abstrak serta menghabiskan banyak waktu, permasalahan tersebut dapat diselesaikan dengan *Topic Modeling* menggunakan metode *Latent Dirichlet Allocation (LDA)* sehingga pengelompokan topik dilakukan secara otomatis dan lebih terstruktur sesuai komposisinya masing-masing.

Kesulitan lain yang sering dialami oleh banyak mahasiswa adalah menentukan tema tugas akhir berdasarkan referensi hasil penelitian sebelumnya (Setijohatmo dkk. 2020). Sebagai contoh pada mahasiswa Prodi Pendidikan Teknik Informatika Dan Komputer sering terjadi kesulitan dalam menentukan tema skripsi, karena kurangnya referensi dari penelitian di Prodi Pendidikan Teknik Informatika Dan Komputer sebelumnya, di mana mahasiswa harus mengunduh di *website repository.unj.ac.id* dan harus mempunyai *account* yang terverifikasi perpustakaan terlebih dahulu. Selain itu terdapat perbedaan pendapat antara substansi dan metodologi yang ingin digunakan oleh mahasiswa dan pembimbing mapun penguji dalam menyelesaikan skripsi (Titin., dkk 2019).

2. Dasar Teori

2.1. Skripsi

Skripsi merupakan kegiatan pembelajaran akhir pada masa studi di perguruan tinggi yang harus diselesaikan oleh setiap mahasiswa agar dapat dinyatakan lulus pada jenjang tertentu. Kegiatan ini merupakan pintu gerbang dari kegiatan akademik sebelumnya yang akan menunjukkan penguasaan keterampilan yang dibutuhkan, pada skripsi mahasiswa menulis artikel ilmiah berdasarkan hasil penelitian yang dilakukan untuk memecahkan masalah ilmiah dengan menggunakan berbagai metode sesuai dengan jenis masalah yang akan dipelajari dan diteliti (Agung dkk., 2020).

2.2. *Data Mining*

Data mining merupakan teknik menemukan sebuah pola dari sekumpulan data yang besar yang tersimpan dalam *database* dengan menggunakan pendekatan statistik dan matematik dalam menghasilkan suatu informasi yang berguna untuk pengguna (Darmawan dkk., 2018). Menurut (Abarca, 2021) *data mining* merupakan teknik ekstraksi data yang sebelumnya tidak diketahui di dalam *database* yang kemudian dibuat sebuah pola data tertentu yang nanti nya dianalisis sesuai karakteristik data sehingga menghasilkan suatu informasi yang berguna untuk saat ini dan juga berikutnya dengan prediksi dari data yang sudah ada dan sudah di proses sebelumnya.

2.3. *Pre-Processing*

Pre-Processing merupakan salah satu tahap dari *data mining* yang digunakan dalam mentransformasikan data menjadi sesuai prosedur dalam *data mining*, dan biasanya akan terjadi perubahan ukuran, hubungan data, normalisasi data, dengan menggunakan beberapa tahap seperti*: case folding, tokenizing,* penghapusan *stopword, lemmatization, dan stemming* (Suad A dkk., 2017).

2.4. *Topic Modeling*

Menurut (Kherwa,2019) *Topic modeling* merupakan metodologi dalam menyajikan data yang berjumlah besar dalam sebuah dokumen menjadi beberapa kata yang menjadi indeks dari dokumen tersebut agar menjadi kumpulan topik-topik yang sesuai dengan isi dokumen, *topic modeling* membantu menemukan pola topik tersembunyi yang ada dalam kumpulan dokumen, memberikan keterangan terhadap isi dokumen dengan cara meringkas, mengatur dan mencari data teks.

2.5. *Latent Dirichlet Allocation (LDA)*

Latent Dirichlet Allocation (LDA) merupakan metode *topic modeling* yang digunakan untuk mengolah dokumen yang berukuran sangat besar, di mana setiap dokumen berisi serangkaian topik tertentu yang sudah diringkas, klasterisasi, dan dihubungkan pada setiap rangkaian kata-kata sehingga menghasilkan daftar topik dengan proporsi yang berbeda-beda di setiap dokumen (Schofield dkk., 2017).

2.6. TF-IDF

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan suatu cara yang digunakan dengan memberikan bobot pada kata terhadap suatu dokumen, di mana gabungan dsari perhitungan bobot disebut dengan *term frequency* (tf) yang berhubungan dengan kemunculan kata pada dokumen tersebut *inverse document frequency* (idf) dari dokumen yang mengandung kata tersebut (Qaiser dkk, 2018; Zhou dkk., 2020).

$$w_{ij} = TF_{IJ} * IDF_j \qquad (1)$$

$$w_{ij} = TF_{IJ} * In(\frac{D}{DF_j}) \quad (2)$$

Keterangan:
i = dokumen ke-d
j = kata ke-I dari kata kunci
W_{ij} = bobot *term* (j) terhadap dokumen (i)

2.7. *Coherence Score*

Coherence score merupakan kumpulan dari kata-kata dari sebuah dokumen yang mempunyai asumsi masing-masing di mana dalam bagian kecil tersebut terdapat topik yang ringkas dan mempunyai korelasi satu sama lain dan *coherence score* digunakan untuk mengukur nilai dari suatu topik yang mempunyai keterkaitan 23tatistic untuk di evaluasi sehingga menghasilkan model topik yang terbaik (Blair dkk.,2020).

$$Cv = S_{set}^{One}\ p_{sw(110)} m_{cos(nlr)\sigma_a} \quad (3)$$

Keterangan
S_{Set}^{One} = segmentasi himpunan pasangan kata W
$P_{Sw(110)}$ = *probabilitas*
$m_{\cos(nlr)}$ = *confirmation measure*
σ_a = agregasi

2.8. *Perplexity*

Perplexity merupakan suatu pengukuran dari model topik yang dihasilkan berdasarkan probabilitas rata-rata dari topik yang terbentuk (Patma, 2021).

$$\text{Per}(D_{test}) = \text{ex}\left\{\frac{\sum Log_p(W_d)}{\sum d\ Nd}\right\} \quad (4)$$

Keterangan :
$P(W_d)$ = peluang dari jumlah kata
N_d = total jumlah kata dalam suatu dokumen ke-d

3. Metodologi

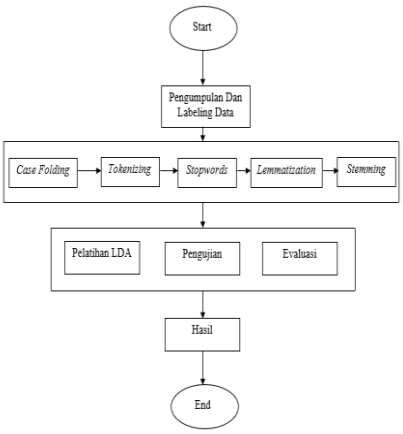
3.1 Alat dan Bahan Penelitian

Alat yang digunakan untuk penelitian ini adalah komputer dengan *library* yang digunakan untuk proses *topic modeling*, visualisasi data, dan analisis data, sedangkan data yang digunakan berasal dari *repository* admin Prodi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta selama 5 tahun terakhir dari tahun lulus 2017-2022 berupa *softcopy*. Berikut penjelasan lengkapnya. Alat -alat yang digunakan dalam melaksanakan penelitian ini, yaitu perangkat keras dan perangkat lunak, antara lain:

1. Laptop dengan *processor* AMD Ryzen 5 2500U dan RAM 8GB
2. *Storage* HDD 1TB dan SSD 256 GB
3. *Python* versi 3.10.2 sebagai Bahasa pemrograman
4. Visual Studio Code sebagai implementasi program
5. *Gensim, Nltk, Sastrawi, PyLDAvis* sebagai *library* untuk pemodelan topik

3.2 Diagram Alir Penelitian

Secara garis besar metode penelitian yang akan dilaksanakan seperti diagram alir berikut ini:



Gambar 3.1 Diagram Alir Penelitian

Pada gambar diatas dijelaskan bahwa untuk mendapatkan suatu pemodelan topik skripsi harus dilakukan dengan melakukan beberapa tahap yaitu: Pengumpulan dan Labeling Data, *Case folding*, *Tokenizing*, *Stopwords*, *Lemmatization*, *Stemming*, dan, Pelatihan dan Evaluasi LDA.

3.3 Perancangan

3.3.1 Labeling data

Tabel 3.1 Proses Labeling Data

Label	Definisi Opsional Labeling
NIM	Nomor Induk Mahasiswa selama melaksanakan perkuliahan.
Tahun Lulus	Tahun lulus mahasiswa pada saat selesai mengerjakan skripsi.
Nama	Nama Mahasiswa yang melakukan penelitian skripsi.
Judul	Judul penelitian yang dilakukan oleh mahasiswa.
Abstrak	Ringkasan singkat dari keseluruhan kandungan dokumen skripsi.
Semester	Semester pada saat mahasiswa selesai mengerjakan skripsi.

3.3.2 *Pre-processing*

1. *Case folding:* Pada tahapan ini data akan diubah menjadi huruf kecil dengan tujuan agar tidak terjadi perbedaan makna jika terdapat penulisan huruf kapital atau tidak.
2. *Tokenizing:* Setelah proses *case folding* dan semua huruf menjadi kecil maka proses selanjutnya adalah membuat kata-kata menjadi potongan tunggal yang berfungsi untuk menghilangkan tanda baca, angka, karakter lain yang tidak ada di dalam alfabet.
3. *Stopwords:* Pada tahap ini akan dilakukan proses penghapusan kata yang dianggap tidak memiliki makna dan paling banyak muncul dalam teks dokumen.
4. *Lemmatization:* Pada proses ini teks akan dinormalisasi dengan mengidentifikasi dan menghapus huruf yang berada di depan dan belakang kalimat sesuai dengan kata dasar yang ada pada kamus.
5. *Stemming*: Pada tahap ini kata yang mempunyai imbuhan diubah menjadi kata dasar, di mana algoritma yang digunakan adalah Algoritma Nazief dan Andriani yang dikombinasikan menggunakan *library* NLTK yaitu sastrawi.

3.3.3 Pemodelan Topik

Pada penelitian ini pemodelan topik menggunakan metode *Latent Dirichlet Allocation* di mana merupakan satu metode yang menghasilkan daftar topik yang mempunyai bobot masing-masing, di mana distribusi yang digunakan adalah *Dirichlet* yang akan memperoleh distribusi topik per dokumen untuk mengalokasikan kata demi kata pada topik yang berbeda (Blei dkk, 2012).

3.3.4 Visualisasi Topik

Visualisasi data menggunakan *PyLDAvis* agar model topik lebih interaktif. Berbasis *website* dengan menafsirkan topik yang berbeda satu sama lain berdasarkan data yang sudah diolah sebelumnya (Onah ,2021). Dan *Word Cloud* agar lebih mudah dalam mempresentasikan pemodelan topik.

3.4 Teknik Analisis Data

3.4.1 TF-IDF

Teknik analisis data dalam penelitian ini menggunakan pembobotan yang telah dilakukan pada tahap *pre-processing* dan mengubah bentuk data ke dalam bentuk numerik dengan metode *Term Frequency (TF*) dengan metode *Frequency Inverse Data (IDF)*.

3.4.2 *Coherence Score*

Coherence score adalah ukuran yang digunakan untuk mengevaluasi *topic modeling* di mana jika skor topik tinggi maka model yang dihasilkan akan baik di mana nantinya akan menjadi acuan untuk membuat model selanjutnya, pada penelitian ini menggunakan limit topik 1-10, 1-20, 1-30.

3.4.3 *Perplexity*

Perplexity merupakan suatu pengukuran dari model topik yang dihasilkan berdasarkan probabilitas rata-rata dari topik yang terbentuk, di mana dilakukan perhitungan dari log teks yang tidak terlihat, semakin kecil nilai *perplexity* maka topik yang dihasilkan semakin baik, pada penelitian ini menggunakan limit topik 1-10, 1-20, 1-30.

4. Hasil dan Analisis

4.1 Deskripsi Hasil Penelitian

Penelitian ini dilakukan untuk mengetahui pemodelan topik skripsi di Prodi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta dengan menggunakan metode *Latent Dirichlet Allocation*, menggunakan 329 dokumen skripsi berupa *soft file* dari tahun lulus 2017-2022. Setelah proses pengumpulan data selanjutnya data diberi label secara manual untuk dilakukan proses *Pre-processing* sehingga menghasilkan 10 topik penelitian diantaranya 50 jenis topik pengembangan media pembelajaran menggunakan multimedia, 42 jenis topik evaluasi pembelajaran, 42 jenis pengembangan *System* informasi berbasis *website*, 36 jenis topik teknik komputer dan Jaringan, 36 jenis topik pengembangan *System* informasi Universitas Negeri Jakarta berbasis *website* menggunakan *user interface, user experience & backend*, 34 jenis topik pengembangan dan evaluasi media pembelajaran *e learning* menggunakan algoritma, 28 jenis topik *data mining*, 26 jenis topik *System requirement* pengembangan perangkat lunak, 14 jenis topik standar pencapaian sumber daya Pendidikan, dan 21 jenis topik evaluasi pengembangan *website*,multimedia, & jaringan. untuk mempermudah komposisi topik skripsi divisualisasikan menggunakan *PyLDAvis* dan *Wordcloud*, serta dilakukan pembobotan TF-IDF untuk mengetahui penyebaran kata dan dievaluasi menggunakan *Coherence Score* dan *Perplexity* untuk mengetahui apakah topik yang dihasilkan sudah baik atau tidak, untuk membuat pemodelan topik selanjutnya.

4.2 Analisis Penelitian

Setelah dilakukan *pre-processing* dan pemodelan topik menggunakan *Latent Dirichlet Allocation (LDA)* selanjutnya topik yang dihasilkan diuji menggunakan TF -IDF, *coherence score, dan perplexity.*

4.2.1 TF-IDF

Tabel 4.1 TF-IDF

Dokumen	D1	D2	D3	D4	D5	..	D325	D326	D327	D328	D329
Akses	0	0	0	0	15,792	..	0	0	0	0	0
Implementasi	0	0	0	2,701	0	..	0	0	0	0	8,105
Video	0	0	18,209	0	0	..	0	0	0	0	0
Wawancara	0	0	0	0	6,264	..	0	0	0	0	6,264
Jaringan	2,652	0	0	0	18,570	..	0	0	0	0	0
:	:	:	:	:	:	..	:	:	:	:	:
Perangkat Lunak	0	0	0	8,837	0	..	0	0	0	0	0
Algoritma	14,876	0	0	0	0	..	0	0	0	0	0
E-Learning	0	0	0	0	0	..	0	0	0	0	0
Testing	0	0	0	0	0	..	0	0	0	5,305	7,958
Machine	16,628	4,157	0	0	0	..	0	0	0	0	0

4.2.2 *Coherence Score*

Tabel 4.2 *Coherence Score Limit 1-10*

Num Topic	*Coherence Score*	*Num Topic*	*Coherence Score*
1	0.2026	6	0.2436
2	0.2175	7	0.2136
3	0.2318	8	0.2578
4	0.2199	9	0.2443
5	0.2429	10	0.2614

Tabel 4.3 *Coherence Score Limit 1-20*

Num Topic	*Coherence Score*	*Num Topic*	*Coherence Score*
1	0.2588	11	0.2423
2	0.2336	12	0.2458
3	0.2535	13	0.2465
4	0.2309	14	0.2427
5	0.2575	15	0.2463
6	0.2458	16	0.2320
7	0.2425	17	0.2436
8	0.2591	18	0.2566
9	0.2536	19	0.2428
10	0.2540	20	0.2357

Tabel 4.4 *Coherence Score Limit 1-30*

Num Topic	*Coherence Score*	*Num Topic*	*Coherence Score*
1	0.2588	16	0.2491
2	0.2499	17	0.2385
3	0.2552	18	0.2512
4	0.2540	19	0.2569
5	0.2332	20	0.2451
6	0.2315	21	0.2440
7	0.2608	22	0.2510
8	0.2275	23	0.2408
9	0.2308	24	0.2387
10	0.2503	25	0.2544
11	0.2475	26	0.2510
12	0.2467	27	0.2391
13	0.2396	28	0.2436
14	0.2415	29	0.2317
15	0.24303	30	0.2483

4.2.3 *Perplexity*

Tabel 4.5 Nilai *Perplexity*

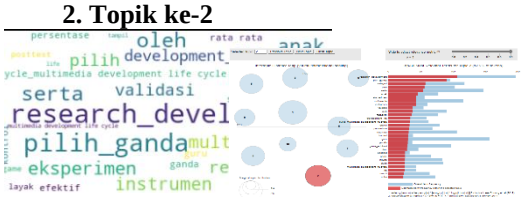
No	**Jumlah Topik**	***Perplexity***
1	10	-5.9895
2	20	-5.9896
3	30	-5.9897

4.3 Pembahasan

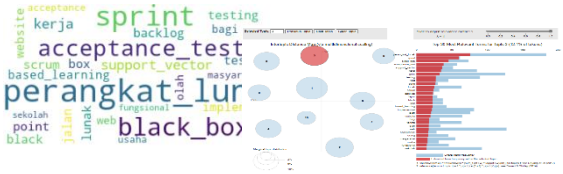
Setelah melakukan penelitian serta pengujian terhadap pemodelan topik skripsi Program Studi Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta menggunakan metode *Latent Dirichlet Allocation* , maka didapatkan hasil evaluasi pengujian dari pemodelan topik yang dihasilkan menggunakan *Coherence Score* dan *Perplexity* sebagai berikut.

1. **Topik ke-1**

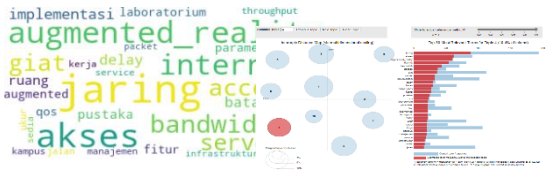
2. **Topik ke-2**



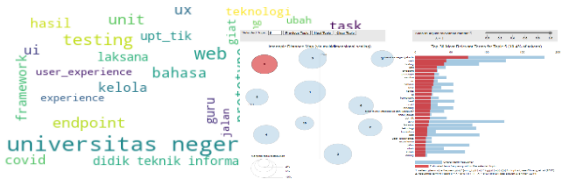
3. **Topik ke-3**



4. **Topik Ke-4**

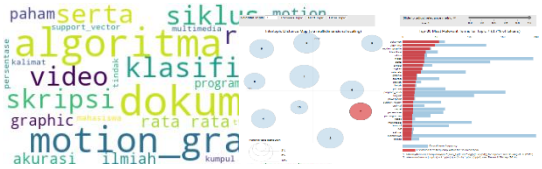


5. **Topik ke- 5**

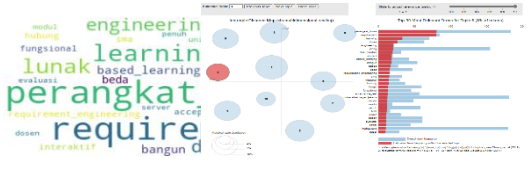


6. **Topik ke-6**

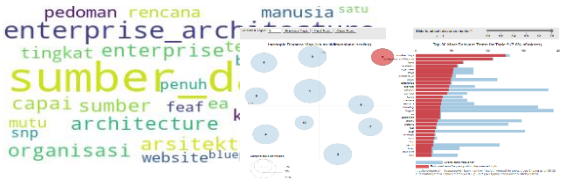
7. **Topik ke-7**



8. **Topik ke-8**



9. **Topik ke-9**



10. **Topik ke-10**

Tabel 4.6 Perhitungan Jumlah Topik

Nomor Topik	Hasil	Jumlah
Topik ke-1	Pengembangan media pembelajaran menggunakan multimedia	50 Dokumen
Topik ke-2	Evaluasi pembelajaran	42 Dokumen
Topik ke-3	Pengembangan *System* informasi berbasis *website*	42 Dokumen
Topik ke-4	Teknik komputer dan jaringan	36 Dokumen
Topik ke-5	Pengembangan *System* informasi Universitas Negeri Jakarta berbasis *website* menggunakan *user interface, user experience, & backend*	36 Dokumen
Topik ke-6	Pengembangan dan evaluasi media pembelajaran *e learning* menggunakan algoritma	34 Dokumen
Topik ke-7	*Data Mining*	28 Dokumen
Topik ke-8	*System requirement* pengembangan perangkat lunak	26 Dokumen

Topik ke-9	Standar pencapaian sumber daya Pendidikan	14 Dokumen
Topik ke-10	Evaluasi pengembangan *website*,multimedia, & jaringan	21 Dokumen
Total		**329 Dokumen**

Tabel 4.7 *Coherence Score* dan *Perplexity*

No	Jumlah Topik	*Coherence Score*	*Perplexity*
1	10	0.2614	-5.9895
2	20	0.2357	-5.9896
3	30	0.2483	-5.9897

Berdasarkan tabel diatas menunjukan bahwa performa topik terbaik berada pada topik 10, karena nilai *coherence score* lebih besar dibandingkan lainnya, dan nilai *perplexity* lebih kecil sehingga dapat disimpulkan nilai topik yang dihasilkan semakin baik.

4.4 Aplikasi Hasil Penelitian

Hasil penelitian ini berupa pemodelan topik skripsi pada Program Studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta menggunakan metode *Latent Dirichlet Allocation* . Hasil evaluasi menggunakan *Coherence Score* dan *Perplexity* menghasilkan 10 topik merupakan pemodelan dengan performa terbaik dengan menghasilkan nilai *Coherence Score* sebesar 0.2614 dan nilai *Perplexity* sebesar -5.9895.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Dalam kurun waktu 5 tahun dihasilkan total 10 topik skripsi Program Studi Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta yang terbagi menjadi 50 jenis topik pengembangan media pembelajaran menggunakan multimedia, 42 jenis topik evaluasi pembelajaran, 42 jenis pengembangan *System* informasi berbasis *website*, 36 jenis topik teknik komputer dan Jaringan, 36 jenis topik pengembangan *System* informasi Universitas Negeri Jakarta berbasis *website* menggunakan *user interface, user experience & backend*, 34 jenis topik pengembangan dan evaluasi media pembelajaran *e learning* menggunakan algoritma, 28 jenis topik *data mining,* 26 jenis topik *System requirement* pengembangan perangkat lunak, 14 jenis topik standar pencapaian sumber daya Pendidikan, dan 21 jenis topik evaluasi pengembangan *website,* multimedia, & jaringan. Evaluasi pemodelan topik dilakukan dengan menggunakan *coherence score* dan *perplexity* di mana 10 topik yang menghasilkan pemodelan yang paling baik dengan nilai *coherence score* 0.2614 dan *perplexity* -5.9895.

5.2 Saran

1. Jumlah data yang digunakan pada penelitian berikutnya bisa lebih banyak lagi agar topik yang dihasilkan lebih baik dan bervariasi.
2. Menggunakan data 3 tahun terakhir, atau tahun-tahun berikutnya agar pemodelan topik yang dihasilkan lebih beragam sesuai dengan kondisi mahasiswa dalam melaksanakan perkuliahan secara offline maupun online.
3. Pada proses *stopwords, normalisasi, dan stemming* Bahasa Indonesia yang digunakan pada *library genism* dan sastrawi sering terdapat kata yang kurang akurat, dan alangkah baiknya dilakukan proses tambahan secara manual pada *term* yang ada pada kode program agar *corpus* yang dihasilkan lebih baik.
4. Visualisasi sebaiknya tidak terlalu banyak mengandung gambar, tabel, dan rumus, untuk mempermudah memahami isi dari topik yang dihasilkan
5. Evaluasi pemodelan yang dilakukan menggunakan akurasi kedekatan dan hasil prediksi dengan *data real* sehingga topik yang dihasilkan dapat diketahui kebenarannya.
6. Perlu dilakukan penelitian lebih lanjut terkait metode yang digunakan untuk mengetahui perbandingan metode yang menghasilkan pemodelan topik yang lebih baik.

Daftar Pustaka:

Abarca, R. M. (2021).. In *Nuevos Systemas de comunicación e información*. http://eprints.poltekkesjogja.ac.id/8119/9.

Anggraini(2020), Latent Dirichlet Allocation untuk pemodelan topik abstrak dokumen skripsi, https://dspace.uii.ac.id/handle/123456789/23556.

Akhmetov, I., Pak, A., Ualiyeva, I., & Gelbukh, A. (2020). *Highly Language-Independent Word Lemmatization Using a Machine-Learning Classifier*. *24*(3), 1353–1364. https://doi.org/10.13053/CyS-24-3-3775.

Alfanzar, A. I., Khalid, K., & Rozas, I. S. (2020). Topic *Modeling* Skripsi Menggunakan Metode Latent Dirichlet Allocation. *JSiI (Jurnal System Informasi)*, *7*(1), 7. https://doi.org/10.30656/jsii.v7i1.2036.

Alhaj, Y. A., Xiang, J., Zhao, D., Al-qaness, M. A. A., Elaziz, M. A. B. D., & Dahou, A. (2019). A Study of the Effects of Stemming Strategies on Arabic Document Classification. *IEEE Access*, *7*, 32664–32671. https://doi.org/10.1109/ACCESS.2019.2903331.

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence*, *50*(1), 138–156. https://doi.org/10.1007/s10489-019-01438-z.

Blei, D., Carin, L., & Dunson, D. (2012). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079.

Chilmi, M. L. C. (2021). Latent dirichlet allocation lda untuk mengetahui topik pembicaraan warganet twitter tentang omnibus law. *Repository.Uinjkt.Ac.Id*. https://*repository*.uinjkt.ac.id/dspace/handle/123456789/56724%0Ahttps://*repository*.uinjkt.ac.id/dspace/bitstream/123456789/56724/1/M. LUVIAN CHISNI CHILMI-FST.pdf.

Darmawan, A., Kustian, N., & Rahayu, W. (2018). Implementasi Data Mining Menggunakan Model SVM untuk Prediksi Kepuasan Pengunjung Taman Tabebuya. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *2*(3), 299. https://doi.org/10.30998/string.v2i3.2439.

Doulaty, M., & Hain, T. (2019). Latent dirichlet allocation based acoustic data selection for automatic speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-Septe*, 3228–3232. https://doi.org/10.21437/Interspeech.2019-179.

Ella. (2020). *Latent Dirichlet Allocation Untuk Pemodelan Tugas Akhir Latent Dirichlet Allocation Untuk Pemodelan*. https://dspace.uii.ac.id/handle/123456789/23556?show=full.

Ervan. (2020). *Analisis Topik Data Tindak Kriminal pada Media Sosial Twitter Menggunakan Metode LDA (Latent Dirichlet Allocation)*. https://dspace.uii.ac.id/handle/123456789/28540%0Ahttps://dspace.uii.ac.id/bitstream/handle/1234567.

F. Holmgren, W., W. Hansen, C., & A. Mikofski, M. (2018). Pvlib Python: a Python Package for Modeling Solar Energy Systems. *Journal of Open Source Software*, *3*(29), 884. https://doi.org/10.21105/joss.00884.

Firdaus, N. (2019). *Buku Ajar*. https://scholar.google.co.id/scholar?hl=id&as_sdt=0%2C5&q=jurnal+artikel+ilmiah&btnG=.

Firman, F. (2018). *Penelitian Kualitatif Dan Kuantitatif*. 1–29. https://doi.org/10.31227/osf.io/4nq5e.

Gabryel, M. (2018). The bag-of-words method with different types of image features and dictionary analysis. *Journal of Universal Computer Science*, *24*(4), 357–371.

Hakim, R. (2020). *Topic Modeling Dokumen Skripsi Menggunakan Metode Latent Semantic*. 1–62.

Han. (2012). Data Mining Concepts and Techniques. In *Nuevos Systemas de comunicación e información*.

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(2), 521–526. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526.

Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2018). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.

Kaur, J., & Kaur Buttar, P. (2018). A Systematic Review on Stopword Removal Algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *April*, 207–210. http://www.ijfrcsce.org.

Kherwa, P., & Bansal, P. (2019). Topic Modeling: A Comprehensive Review EAI Endorsed Transactions on Scalable Information Systems. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 1–16.

MAHMUD, S. (2019). *Penggunaan Pemodelan Topik Untuk Mengetahui Perkembangan Minat Mahasiswa Akan Topik Tugas Akhir Menggunakan Metode …*. 2017. http://eprints.umg.ac.id/848/.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., Delen, D., 2012. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Academic Press/Elsevier,

Onah, D., & Pang, E. (2021). Mooc Design Principles: Topic *Modeling*-Pyldavis Visualization & Summarization of Learners' Engagement. *EDULEARN21 Proceedings*, *1*(July), 1082–1091. https://doi.org/10.21125/edulearn.2021.0282.

Panda, M. (2018). Developing an Efficient Text Pre-Processing Method with Sparse Generative Naive Bayes for Text Mining. *International Journal of Modern Education and Computer Science*, *10*(9), 11–19. https://doi.org/10.5815/ijmecs.2018.09.02.

Patma. (2021). Analisis Topik *Modeling* Terhadap Penggunaan Sosial Media Twitter Oleh Pejabat Negara. : https://www.researchgate.net/publication/357700245.

Ridha, N. (2017). Proses Penelitian, Masalah, Variabel, dan Paradigma Penelitian. *Jurnal Hikmah*, *14*(1), 62–70.

http://jurnalhikmah.staisumatera-medan.ac.id/index.php/hikmah/article/download/10/13.
Ririd, A., Saputra, P. Y., & ... (2019). *System* koreksi kesalahan pengetikan kata kunci dalam pencarian artikel menggunakan algoritma jaro-winkler. *Seminar Informatika …*, 60–65. http://jurnalti.polinema.ac.id/index.php/SIAP/article/view/368/159.
Röder, M. (2015). exploring the Space of Topic Coherence Measures http://dx.doi.org/10.1145/2684822.2685324.
Roifa, A. N. (2018). *Text Mining Dengan Metode Naive Bayes Classifier Untuk Mengklasifikasikan Berita Berdasarkan Konten*. https://*repository*.its.ac.id/51007/.
(Rosid, 2020). Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi doi: 10.1088 /1757-899X/874/1/012017, Materials Science and Engineering 874 (2020) 012017
Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Pre-Processing for Latent Dirichlet Allocation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 432–436.
Setijohatmo, U. T., Rachmat, S., Susilawati, T., Rahman, Y., & Kunci, K. (2020). Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik. *Prosiding The 11th Industrial Research Workshop and National Seminar*, 402–408.
Suad A. Alasadi, & Wesam S. Bhaya. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107.
Template Riwayah. (2020). DOI: http://dx.doi.org/10.21043/riawayh.v5i2.4933
Titin Supiani., *et al*.(2019). *Buku panduan skripsi mahasiswa*. *Edisi 2019 https://ft.unj.ac.id/*.
Wahyudi, D., Susyanto, T., & Nugroho, D. (2017). Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia. *Jurnal Ilmiah SINUS*, *15*(2), 49–56. https://doi.org/10.30646/sinus.v15i2.305.
Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications, 181(1), 25–29.
Yoren.(2018). Perbandingan Raw Tf Dan Binary Tf Pada *System* Pencarian Di Situs Museum Wayang the Comparison of Raw Tf and Binary Tf on Searching System in *Website* of Kekayon Museum of Puppets inYogyakarta. https://*repository*.usd.ac.id/32223/.
Zhou, T., Wang, Y., & Zheng, X. (2020, December). C*hinese text classification method using FastText and term frequency-inverse document frequency optimization*. In Journal of Physics: Conference Series (Vol. 1693, No. 1, p. 012121). IOP Publishing.