Technologica

Vol.5(1): Hal 125-135 (Januari 2026)

ISSN: 2827-9492
(*Online*)

Pemodelan topik Dokumen Tesis menggunakan Metode *Latent dirichlet allocation*

Topic Modeling of Thesis Documents with Latent dirichlet allocation

Mardiah[*,1], Anis Fitri Nur Masruriyah[2], Ade Hikma Tiana[3], Bobby Suryo Prakoso[4], Rizky Tito Prasetyo[5], Sanggi Bayu Ardika[6]

[1,3,4,5]Program Studi Sistem Informasi, Universitas Pembangunan Nasional Veteran Jakarta
[2,6]Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta

*Penulis Korespondesi
Email: mardiah@upnvj.ac.id

Abstrak. Penelitian merupakan suatu langkah yang dilakukan untuk mengembangkan ilmu pengetahuan dan mencari kebenaran. Dasar dalam melakukan penelitian adalah membaca dokumen penelitian sebelumnya. Namun, pencarian dokumen penelitian yang saling berhubungan seringkali membutuhkan banyak waktu. Maka, dibutuhkan pemodelan topik yang dapat mengelompokkan dokumen berdasarkan topiknya agar membantu peneliti dalam melaksanakan tugasnya. Penelitian ini menggunakan metode *Latent dirichlet allocation* untuk memodelkan topik dalam dokumen, dan menggunakan perhitungan nilai coherence untuk menentukan jumlah topik yang akan dimodelkan. Hasil analisis menunjukkan bahwa pemodelan topik dengan metode *Latent dirichlet allocation* berhasil membagi 340 dokumen tesis dalam 5 topik utama dengan nilai coherence yaitu 0.445. Karakteristik yang terdapat dalam tiap topik merupakan bidang kajian tertentu yaitu sistem informasi, kebakaran lahan, bioinformatika, pengolahan citra, dan robotika. Hasil yang didapatkan menunjukkan metode LDA telah berhasil mengelompokkan dokumen dalam kesamaan topik atau kajian tertentu.
Kata kunci: *latent dirichlet allocation*, pemodelan topik, coherence score, dokumen tesis

Abstract. *Research is an essential step in the development of science and the pursuit of truth. One of the fundamental stages in conducting research is reviewing previous scholarly documents. However, finding related research documents often requires considerable time and effort. Therefore, topic modeling is needed to group documents based on their topics, helping researchers carry out their work more efficiently. This study employs the Latent dirichlet allocation (LDA) method to model topics within documents and uses the coherence score to de*termi*ne the optimal number of topics. The dataset used in this study consists of 340 thesis documents collected from the IPB University Repository between 2017 and 2020. The results show that the documents were successfully grouped into five distinct topics with a coherence score of 0.445, indicating that the documents were effectively categorized based on their content.*
Keywords: *latent dirichlet allocation, pemodelan topik, coherence score, thesis document*

1. Pendahuluan

Text mining adalah proses ekstraksi informasi berharga dari data teks yang tidak terstruktur, dengan tahapan mulai dari pengumpulan data, *preprocessing*, pemodelan, hingga analisis. Dalam prosesnya, *text mining* melibatkan identifikasi pola, tren, dan pengetahuan yang relevan dari data

teks tidak terstruktur melalui metode komputasi dan linguistik (Sabna, 2020). Salah satu tujuan *text mining* adalah memahami isi atau pokok bahasan teks dalam jumlah besar, *text mining* sudah digunakan dalam berbagai bidang, seperti bidang kesehatan dalam penelitian (Ahmad, Shah, & Lee, 2023), pendidikan (Ahadi, Singh, Bower, & Garrett, 2022) dan *business intelligence* (Li et al., 2022). Pemanfaatan *text mining* dalam pendidikan tinggi dapat meningkatkan kemampuan institusi untuk menganalisis dan mengoptimalkan berbagai aspek ekosistem akademik, mulai dari evaluasi kurikulum hingga peningkatan kualitas pembelajaran (Cahyanto, Pamungkas, & Zulkarnain, 2024) Teknik ini memungkinkan identifikasi pola-pola tersembunyi dalam data tekstual yang melimpah, seperti transkrip perkuliahan, makalah penelitian, dan umpan balik mahasiswa, yang sebelumnya sulit diakses melalui metode analisis tradisional (Cahyani & Arif, n.d.). Pemanfaatan teks mining pada dokumen penelitian sudah dilakukan dalam (Pham et al., 2021) untuk membuat *workflow* yang bisa mempercepat dan mempermudah peneliti dalam menyeleksi abstrak penelitian yang relevan untuk systematic review. Melalui sytematic review, peneliti dapat mengidentifikasi literatur atau penelitians ebelumnya, mengevaluasi metodologi, dan mensintesis temuan dari berbagai studi untuk memperoleh pemahaman komprehensif tentang suatu topik (Nurhidayah, 2024)

Topik dalam penelitian merupakan suatu hal yang harus ditentukan dengan dasar jurnal atau publikasi penelitian sebelumnya. Dokumen penelitian biasanya tersimpan *online* dalam repositori kampus atau dalam *repository publisher*. Untuk mencari topik yang sesuai dengan penelitian yang sedang atau akan dijalani, peneliti harus membaca dokumen yang saling terkait dengan topik penelitiannya. Pengelompokkan dokumen sesuai dengan topiknya akan sangat bermanfaat bagi peneliti yang sedang mencari referensi untuk penelitian. Pendekatan yang dapat digunakan untuk mengatasi masalah ini adalah ekstraksi informasi dari dokumen dengan *text mining* dan pemodelan topik.

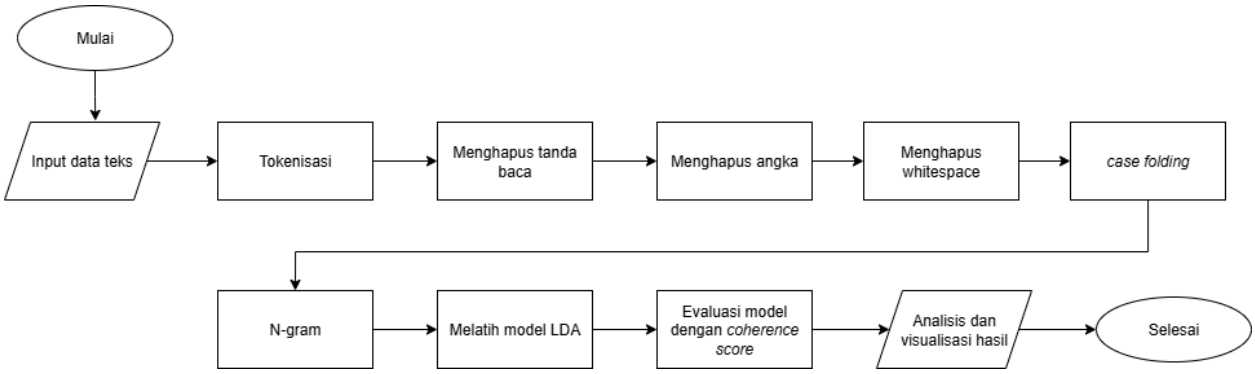
Penelitian ini akan menggunakan dokumen tesis *fulltext* dari judul, abstrak, sampai dengan kesimpulan. Data teks bersumber dari repository IPB yaitu dokumen tesis dari tahun 2017 sampai dengan 2020 sebanyak 340 dokumen. Dalam *text mining*, ada satu proses penting yaitu praproses data teks. Perbedaan pada karakteristik dokumen seperti bahasa dan ukuran dokumen berdasarkan jumlah kata dapat menjadikan perbedaan kebutuhan untuk proses praproses pada data teks (Hickman, Thapa, Tay, Cao, & Srinivasan, 2022) Tujuan praproses data adalah untuk menghapus bagian yang tidak diperlukan agar data siap untuk di latih modelnya menggunakan metode topic modelling. *Topic modelling* adalah salah satu teknik dalam *text mining* yang digunakan untuk mengidentifikasi tema atau topik tersembunyi dari kumpulan dokumen tanpa perlu membaca tiap dokumen.

Terdapat banyak metode dalam pemodelan topik antara lain *Latent Semantic Indexing* (LSI) (Damayanti, Purwitasari, & Suciati, 2021), *Non Negative Factorization* (NMF) dan N-gram (Afidh & Syahrial, 2023) dan *Latent dirichlet allocation* (LDA). Pada penelitian ini, akan dibatasi hanya menggunakan LDA saja. LDA adalah model probabilistik generatif yang bertujuan untuk menemukan struktur tersembunyi dalam kumpulan dokumen atau korpus yang pertama kali diusulkan dalam penelitian (Blei, Ng, & Edu, 2003) . LDA mengasumsikan bahwa setiap dokumen tersusun atas berbagai topik, dan setiap topik terdistribusi dalam probabilitas kata. Dengan mengimplementasikan LDA ke korpus teks, kita dapat secara otomatis mengidentifikasi topik yang terdapat dalam teks tersebut, memahami hubungan antar dokumen. Pemodelan topik menggunakan LDA dengan dokumen berbahasa indonesia sudah dilakukan pada penelitian (Uray

Nur Khadijah & Nuri Cahyono, 2024) yang sudah memodelkan data teks pariwisata yogyakarta dari twitter sejumlah 7758 baris. Hasil penelitian tersebut menunjukkan LDA telah berhasil memodelkan topik dengan dokumen berbahasa indonesia dengan membagi topik = 3 topik yang berbeda. Pemanfaatan LDA pada penelitian Salah satu langkah utama dalam pemodelan topik adalah menentukan jumlah topik yang yang akan dimodelkan. Terlalu sedikit topik dapat menyebabkan generalisasi yang berlebihan dan hilangnya detail penting, sementara terlalu banyak topik dapat menghasilkan topik yang redundan atau tidak dapat diinterpretasikan. Penelitian - (Uray Nur Khadijah & Nuri Cahyono, 2024) menggunakan nilai coherence untuk menentukan jumlah topik yang akan dimodelkan dalam LDA. Penelitian ini akan melakukan hal yang sama dalam evaluasi dan pemodelan, perbedaan terdapat pada dataset yang digunakan.

2. Metode

Metode yang digunakan dalam penelitian ini ditunjukkan pada gambar 2.1 dimulai dengan input dokumen sampai dengan evaluasi model LDA dengan coherence score.



Gambar 1 Metode Penelitian

2.1 Input Data Teks

Data teks yang digunakan dalam penelitian ini sebagai input utama diambil dari repository IPB university. Data teks berupa dokumen tesis magister ilmu komputer dari tahun 2007 sampai dengan 2020 berjumlah 340 dokumen. Teks ini merujuk pada penelitian yang telah dilakukan dalam (Mardiah, Annisa, & Nidya Neyman, n.d.)

2.2 Praproses Teks

Praproses teks adalah suatu langkah dalam mengubah atau menghapus elemen dari teks asli dan hasil dari praproses adalah teks yang diinginkan. Pada penelitian ini, dilakukan 5 langkah praproses data yaitu tokenisasi, menghapus tanda baca, menghapus angka, menghapus *whitespaces,* dan *case folding*. Penjelasan mengenai tiap langkah praproses teks akan menggunakan kalimat “Sistem informasi digunakan dalam pengambilan keputusan manajerial Nomor 1!." Sebagai contoh guna memperjelas dan mempermudah pemahaman untuk tiap tahapan praproses teks.

1. Tokenisasi

   Memisahkan kalimat menjadi kata per kata.

   ['Sistem', 'informasi', 'digunakan', 'dalam', 'pengambilan', 'keputusan', ' manajerial', ‘Nomor’,’1’]

2. Menghapus Tanda Baca & Angka:
   [‘Sistem’, ‘informasi’, ’digunakan’, ‘dalam’, ‘pengambilan’, ‘keputusan’,’ manajerial’, ’Nomor’,’1’,’.’]
3. Case folding
   Mengubah huruf kapital menjadi huruf kecil
   [‘sistem’, ‘informasi’, ’digunakan’, ‘dalam’, ‘pengambilan’, ‘keputusan’,’ manajerial’, ’nomor’,’1’,’.’]
4. Menghapus Whitespace:
   ['sistem', 'informasi', 'digunakan', 'dalam', 'pengambilan', 'keputusan', 'manajerial', ‘nomor’]
5. Menghapus *stopwords*
   Proses untuk menghapus *“stopwords”* atau kata yang tidak diperlukan dalam dokumen.
   ['sistem', 'informasi', 'pengambilan', 'keputusan', 'manajerial']

2.3 N-gram

N-gram merupakan suatu deret kata yang berjumlah n kata. n=2 (2-gram) atau dikenal dengan bigram merupakan suatu susunan kata yang terdiri dari dua kata. n=3 (3-gram) atau dikenal dengan trigram merupakan suatu susunan kata yang terdiri dari tiga kata. Pda penelitian ini, n gram jan menggunakan jumlah n = 2 atau susunan bigram saja.

2.4 Melatih model LDA

Latent dirichlet allocation atau LDA adalah suatu metode yang dapat digunakan untuk pemodelan topik. Berikut langkah-langkah dalam LDA:

1. Preprocessing: Membersihkan data teks agar siap untuk dimodelkan.
2. Vectorisasi: ubah teks ke format numerik, pada penelitian ini dengan *Bag of Word* (Bow)

2.5 Evaluasi dengan Coherence Score

Menentukan jumlah topik pada LDA merupakan satu hal yang dapat dievaluasi, salah satunya dengan coherence score. Evaluasi dengan coherence score bertujuan untuk mengetahui seberapa besar hubungan semantik antar topik yang ada dalam dokumen.

2.6 Analisis dan visualisasi hasil

Pada penelitian ini, digunakan library *pyLDAvis* untuk melakukan visualisasi hasil pemodelan topik dengan LDA. Visualisasi ini bertujuan untuk mempermudah interpretasi topik-topik yang terbentuk serta melihat distribusi dan hubungan antar topik dalam korpus dokumen. Setiap topik direpresentasikan dalam bentuk gelembung (bubble) pada bidang dua dimensi, di mana ukuran gelembung menunjukkan proporsi dokumen yang tergolong dalam topik tersebut, dan jarak antar gelembung merepresentasikan kemiripan semantik antar topik. Selain itu, *pyLDAvis* juga menampilkan daftar term/kata utama dalam setiap topik beserta frekuensinya, sehingga memudahkan peneliti dalam memahami karakteristik tiap topik yang terbentuk dari hasil pelatihan model.

3. Hasil dan Pembahasan

3.1 Hasil Penelitian

Deskripsikan hasil setiap tahapan berikut ini:

1. Input Data Teks,
   Jelaskan sumber datanya, seperti apa data yg diperoleh/siap untuk dilakukan pra-proses.

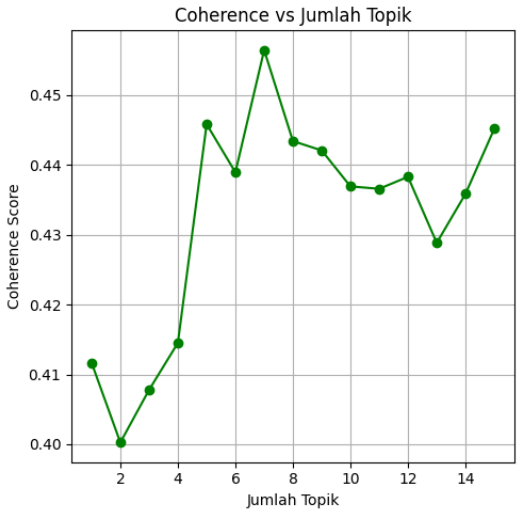
2. Praproses Teks,
   Data yang digunakan dalam penelitian ini diambil dari https://repository.ipb.ac.id/ berupa dokumen tesis magister ilmu komputer dari tahun 2007 sampai dengan 2020 berjumlah 340 dokumen
3. N-gram
   Analisis N-gram bertujuan untuk mengidentifikasi pola kemunculan kata dalam dokumen. N-gram merupakan metode pemodelan teks yang membagi kalimat menjadi rangkaian berurutan yang terdiri sejumlah *n* kata. Pada penelitian ini diterapkan tiga jenis N-gram yaitu unigram (1-gram), bigram (2-gram), dan trigram (3-gram). Hasil dari 1-gram terdapat 42677 kata, 2-gram 330477 kata, dan 3-gram 667398 kata. Top-10 kata dengan frekuensi tertinggi ditunjukkan pada Tabel. 1 Unigram menganalisis kata tunggal tanpa memperhatikan konteks kata lain sebelum atau sesudahnya. Unigram mampu menggambarkan kata yang dominan, namun tidak dapat menjelaskan hubungan antar kata sehingga konteks makna seringkali hilang. Hasil top-10 Bigram menunjukkan pasangan kata yang muncul secara berurutan. Informasi yang didapatkan menunjukkan pola kata yang merepresentasikan konteks tertentu. Berdasarkan hasil tersebut bigram lebih relevan. Trigram memodelkan tiga kata yang berurutan, hasil top-10 trigram pada penelitian ini menunjukkan perulangan unigram yang maknanya cenderung tidak jelas. Maka, pada penelitian ini bigram dipilih sebagai model yang paling sesuai dengan karakteristik data.

Table 1 Hasil n-gram

1-gram	2-gram	3-gram
game	bagus bagus	bagus bagus bagus
bagus	dark sistem	mantap mantap mantap
main	game seru	keren keren keren
seru	mantap mantap	game seru game
tim	game rusak	seru game seru
sistem	keren keren	honor of king
dark	game bagus	banget bagus banget
keren	seru game	bagus banget bagus
aja	bagus banget	game rusak game
banget	honor of	honor of kings

4. Melatih model LDA
   Pada tahap ini dilakukan pelatihan model LDA untuk mengetahui struktur topik yang terdapat pada 340 dokumen tesis. Pelatihan tidak dilakukan hanya sekali, melainkan melalui beberapa percobaan dengan variasi jumlah topik, yaitu k = 4, 5, 6 dan 7. Setiap model yang dihasilkan kemudian dievaluasi menggunakan *coherence score* untuk mengetahui seberapa baik hubungan antar kata dalam masing-masing topik.
5. Evaluasi dengan Coherence Score
   Tahap evaluasi dengan coherence score dilakukan dengna tujuan untuk untuk menilai kualitas topik yang dihasilkan oleh model LDA. Berikut ditunjukkan pada gambar 2 merupakan grafik nilai coherence. Nilai coherence paling tinggi adalah saat jumlah topic = 7 yaitu 0.456, namun setelah dilakukan train model dan visualisasi dengan pyLDAvis, masih terdapat topik yang tumpang tindih. Maka dilakukan train ulang dengan menurunkan jumlah topik. Nilai coherence score saat jumlah topik = 5 adalah 0.445. Jika ditinjau dari selisih

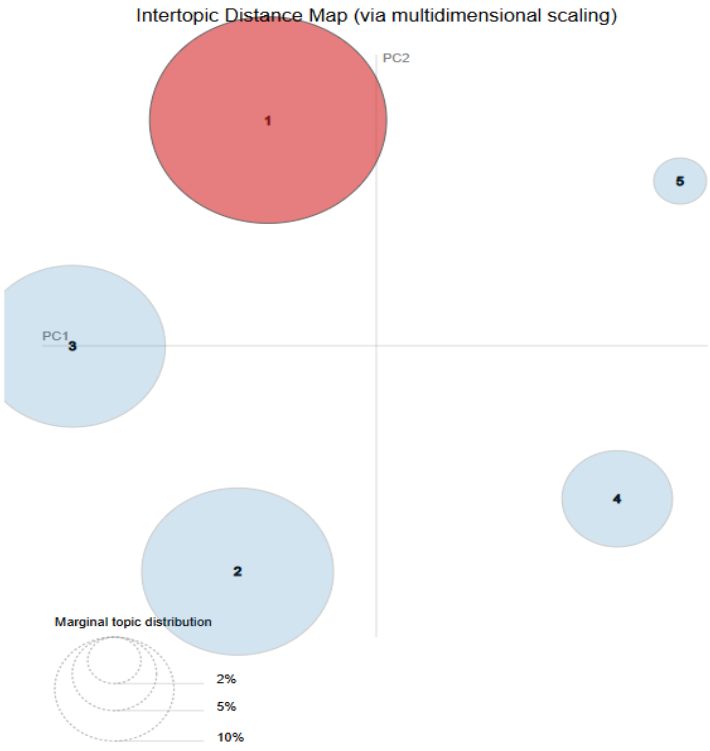
nilai coherence, untuk jumlah topik 5 dan 7 tidak terlalu banyak selisihnya yaitu hanya 0.011. Nilai coherence score 0.445 masih dapat dimaknai nilai yang cukup tinggi.



Gambar 2 Grafik nilai coherence

6. Analisis dan visualisasi hasil

   Saat dilakukan visualisasi dengan jumlah topik = 5, jarak antar topik baru mendapatkan hasil lebih baik karena tidak lagi terdapat topik yang tumpang tindih dan juga jarak gelembung tiap topik terlihat jelas seperti ditunjukkan ada gambar 3. Topik 1 yang ditandai dengan gelembung warna merah memiliki ukuran paling besar, yang dapat dimaknai bahwa dokumen yang tergabung dalam topik 1 memiliki jumlah paling banyak jika dibandingkan dengan dokumen yang derdapat dalam topik lainnya (2 sampa dengan 5).

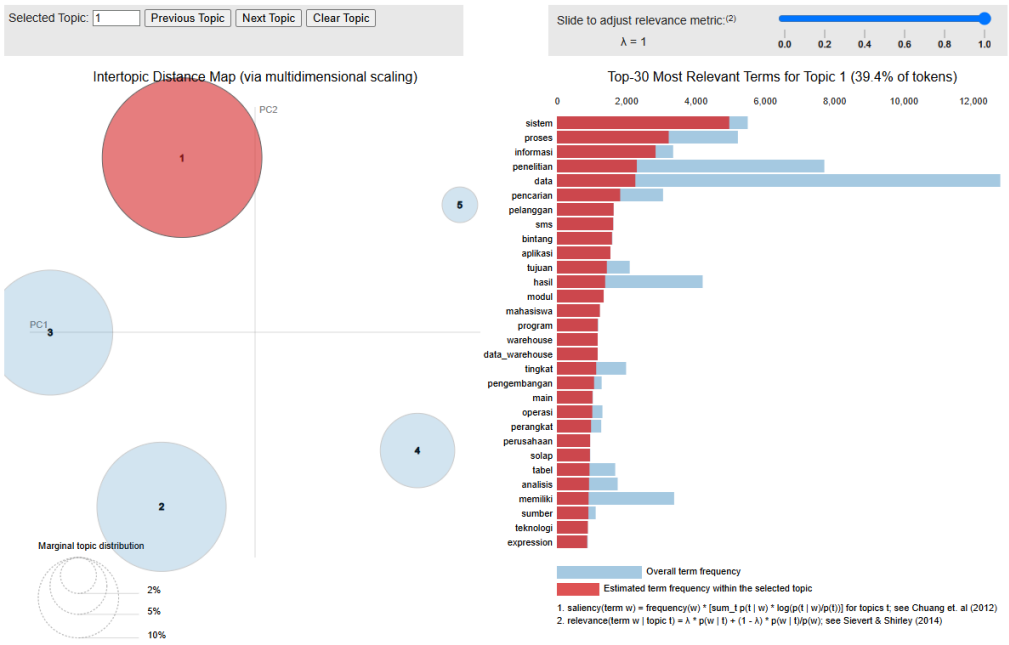


Gambar 3 Visualisasi jumlah topik = 5 dengan pyLDAvis

Topik 1

Visualisasi term topik 1 ditunjukkan pada gambar 4. Untuk semua visualisasi topik *bar* warna merah menunjukkan estimasi frekuensi kata tersebut dalam topik yang dipilih, *bar* warna biru menunjukkan Frekuensi kata tersebut di seluruh korpus (semua topik). masing-masing term atau kata yang terdapat dalam topik 1 memiliki estimasi frekuensi yang cukup tinggi jika dibandingkan dengan bar biru yang merupakan frekuensi kata tersebut dalam semua dokumen.

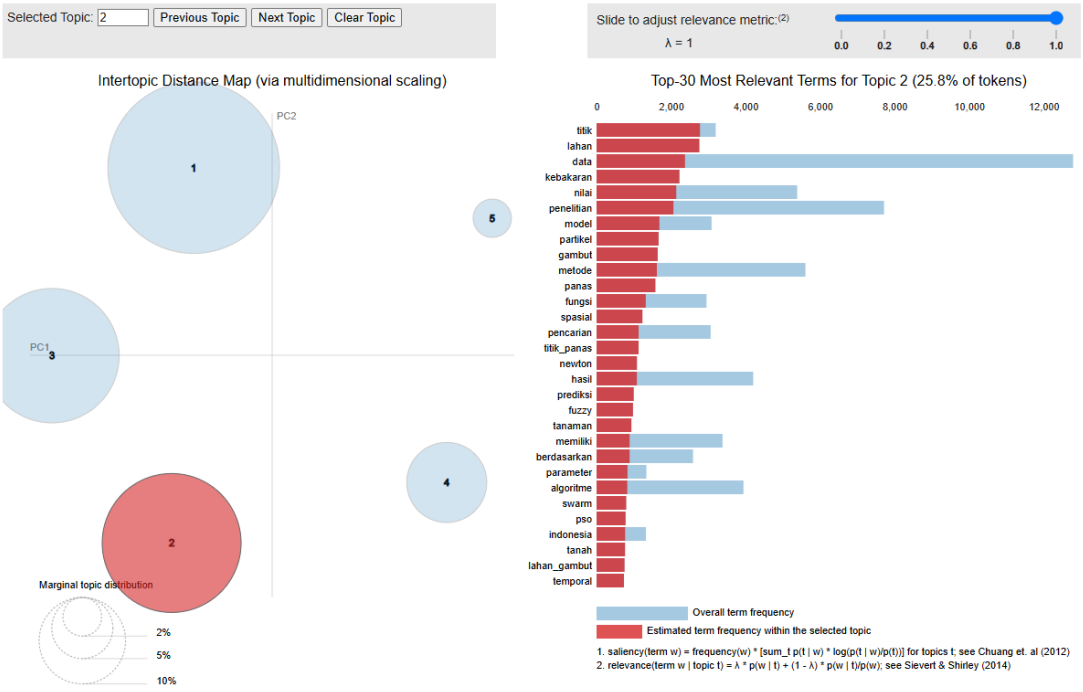
Top -10 term yang terdapat dalam topik 1 mencakup sekitar 39,4% dari keseluruhan term yaitusistem, proses, informasi, penelitian, data, pencarian, pelanggan, sms, bintang, aplikasi. Berdasarkan term dari topik 1 ini, kita dapat menyimpulkan bahwa dokumen yang terdapat dalam topik 1 membahas penelitian yang membahas tentang aplikasi, data dan pelanggan. Topik ini berpusat pada sistem informasi, terutama terkait proses pengolahan data, pencarian informasi, dan aplikasi teknologi dalam pengembangan sistem atau layanan (seperti SMS, pelanggan, SOLAP, *data warehouse*). Kemungkinan besar mencakup tema tentang pengembangan dan evaluasi sistem informasi atau *decision support system* (DSS).



Gambar 4 Visualisasi topik 1

Topik 2

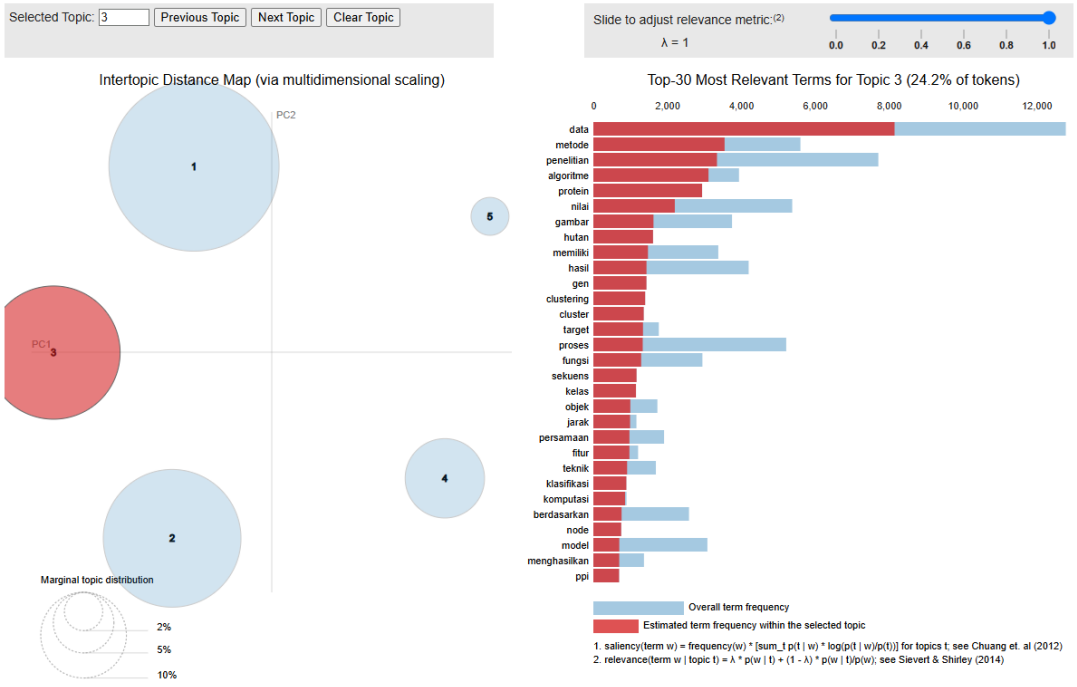
Beberapa term pada topik 2 yang ditunjukkan pada gambar 5 adalah titik, lahan, data, kebakaran, lahan gambut, pso, titik panas, fuzzy. Berdasarkan term yang terdapat pada topik 2. Term tersebut mencakup 25.8% relevansi topik kita dapat menyimpulkan dokumen yang terdapat dalam topik 2 adalah dokumen yang membahas tentang pemodelan kebakaran lahan gambut, termasuk analisis spasial dan temporal, prediksi titik panas, serta penggunaan metode seperti algoritma PSO, fuzzy.



Gambar 5 Visualisasi topik 2

Topik 3

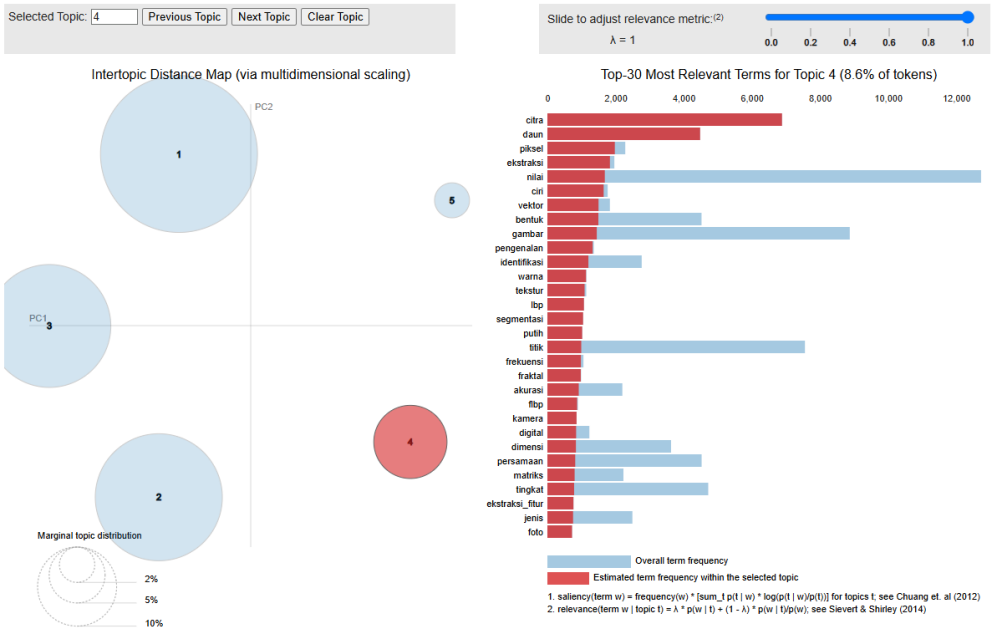
Dokumen yang terdapat dalam topik 3 merupakan dokumen yang membahas gen, *clustering*, sekuens, protein ditunjukkan pada gambar 6. Topik 3 ini membahas mengenai bioinformatika atau data sains berbasis biologi, fokusnya pada metode komputasi untuk analisis data genetik/protein seperti clustering, klasifikasi, dan pengolahan data sekuens/protein.



Gambar 6 Visualisasi topik 3

Topik 4

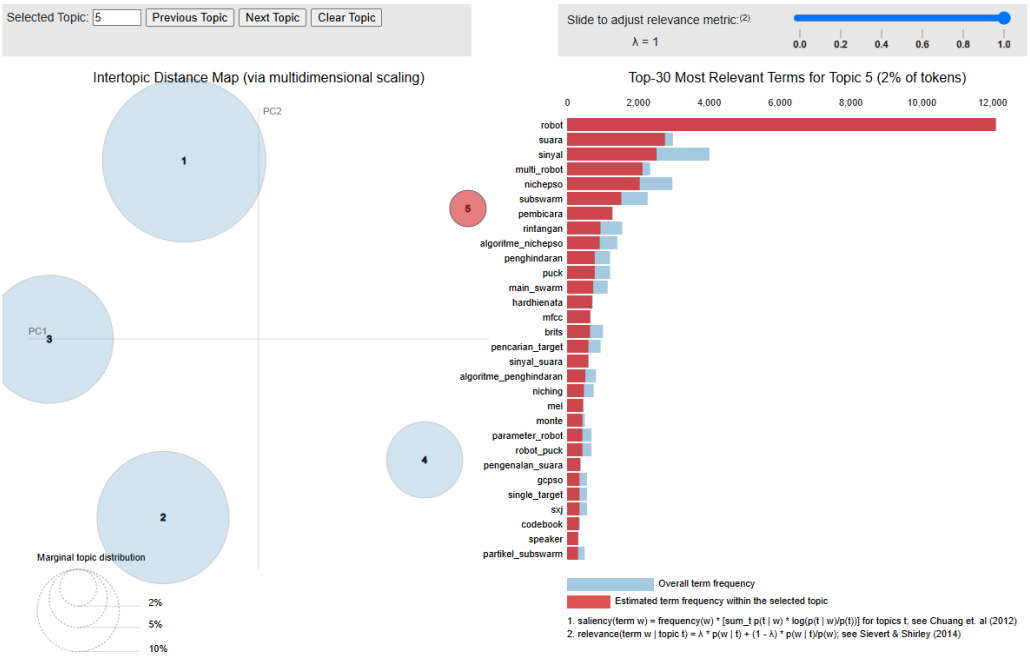
Dokumen yang terdapat pada topik 4 merupakan dokumen yang membahas tentang citra, daun. Visualisasi topik 4 ditunjukkan pada gambar 7. Topik 4 membahas mengenai pengolahan citra digital, khususnya terkait citra daun atau objek visual lainnya. Fokus pada teknik ekstraksi fitur (LBP, warna, tekstur), segmentasi citra, dan klasifikasi visual berdasarkan fitur gambar. Relevan untuk tema computer vision atau klasifikasi citra.



Gambar 7 Visualisasi topik 4

Topik 5

Visualisasi topik 5 ditunjukkan pada gambar 8. Dokumen yang terapat pada topik 5 memiliki jumlah yang paling sedikit. Topik 5 membahas tentang robot, sara, sinyal, dan multi_robot. Topik 5 memiliki persentase *related* term terkecil diantara topik lainnya yaitu 2%.



Gambar 8 Visualisasi topik 7

Topik 1 sampai dengan 5 memiliki kajian topik yang berbeda, serta jumlah dokumen yang berbeda di setiap topik. Penelitian ini dapat mempermudah peneliti untuk mencari dokumen dengan kemiripan topik yang sesuai dengan kajian penelitian.

3.2 Pembahasan

Penelitian ini menghasilkan 5 topik dari pemodelan dengan metode *Latent dirichlet allocation* dengan nilai coherence 0.445, nilai ini menunjukkan adanya keterkaitan yang moderat atau tidak terlalu tinggi diantara term yang membentuk topik didukung oleh riset (Candra Mahatagandha, Agus Sanjaya, & Selatan, 2022) Pada visualisasi tergambar 5 lingkaran yang mana tiap satu lingkaran mewakili satu topik. Perbedaan ukuran lingkaran menunjukkan presentse kata yang terdapat dalam topik tersebut. Tiap lingkaran yang tidak tumpang tindih menunjukkan tiap topik membahas konsep, ide, atau tema yang secara signifikan berbeda satu sama lain

Hasil penelitian ini seiring dengan berbagai penelitian (Uray Nur Khadijah & Nuri Cahyono, 2024) yang telah membuktikan efektivitas LDA dalam memodelkan dokumen berbahasa Indonesia. Selain itu, relevan pada penelitian (Sabna, 2020) yang menyatakan keberhasilan pemodelan topik sangat dipengaruhi oleh karakteristik data serta tahapan praproses yang diterapkan.

Penelitian ini bermanfaat bagi perkembangan IT, khususnya pada bidang *text mining*, karena menunjukkan bagaimana pemodelan topik dapat digunakan untuk mengelompokkan dokumen dan mempermudah proses analisis teks dalam jumlah besar. Pendekatan ini mendukung pengembangan sistem pengelolaan pengetahuan, pencarian dokumen yang lebih baik lagi, serta penerapan topic modelling khususnya *Latent dirichlet allocation* pada korpus berbahasa Indonesia.

4. Kesimpulan

Kesimpulan dari penelitian ini adalah *Topic modelling* dengan metode *Latent dirichlet allocation* telah berhasil dilakukan, dokumen yang terbagi berdasarkan kemiripan topik kajian dapat mempermudah peneliti dalam mencari dokumen yang sesuai dengan topik kajiannya. Penelitian ini dilakukan menggunakan data 340 dokumen tesis yang tersusun dari judul sampai dengan kesimpulan.

Pemodelan topik dibagi kedalam 5 topik yang berbeda berdasarkan hasil train model LDA dan juga evaluasi dengan nilai coherence. Topik dengan jumlah 5 ditetapkan karena memiliki visualisasi yang paling bersih jika dibandingkan dengan topik berjumlah 7 yang memiliki nilai coherence paling tinggi namun setelah divisualisasi memiliki topik yang tumpang tindih.

Hasil dari penelitian ini adalah topik dari data tersebar dengan baik, Hal ini ditunjukkan dengan jarak antara lingkaran topik cukup juah. Term yang berasal dari tiap topik juga terlihat bahwa tiap topik memiliki term yang bervariasi, hal ini menunjukkan bahwa tiap dokumen yang tersusun dalam topik yang berbeda, membahas topiknya masing masing. Sesuai dengan keanggotaan dokumen tersebut terhadap topiknya.

Daftar Pustaka

Afidh, R. P. F., & Syahrial. (2023). Pemodelan Topik Menggunakan n-Gram dan Non-negative Matrix Factorization. *Jurnal Informasi Dan Teknologi*, 265–275. https://doi.org/10.60083/jidt.v5i1.385

Ahadi, A., Singh, A., Bower, M., & Garrett, M. (2022, March 1). *Text mining* in Education—A Bibliometrics-Based Systematic Review. *Education Sciences*, Vol. 12. MDPI. https://doi.org/10.3390/educsci12030210

Ahmad, P. N., Shah, A. M., & Lee, K. Y. (2023, May 1). A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain. *Healthcare (Switzerland)*, Vol. 11. MDPI. https://doi.org/10.3390/healthcare11091268

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). *Latent dirichlet allocation* Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

Cahyani, L., & Arif, M. (n.d.). *Text mining* untuk Pengelompokan Skripsi di Prodi Pendidikan Informatika Universitas Trunojoyo Madura. In *Jurnal Ilmiah Edutic* (Vol. 8).

Cahyanto, H. N., Pamungkas, P., & Zulkarnain, O. (2024). *Pengaruh Penggunaan Chatgpt Terhadap Kemandirian Mahasiswa Dalam Menyelesaikan Tugas Akademik*. *8*(1).

Candra Mahatagandha, P., Agus Sanjaya, N. E., & Selatan, K. (2022). Pemodelan Topik Skripsi Menggunakan Metode Lda. In *Jnatia* (Vol. 1).

Damayanti, P., Purwitasari, D., & Suciati, N. (2021). Eliminasi Data Non-topic menggunakan Pemodelan Topik untuk Peringkasan Otomatis Data Tweet dengan Konteks Covid-19. *urnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, *8*, 199–08.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for *Text mining* in Organizational Research: Review and Recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., & Yu, Z. (2022, October 1). *Text mining* of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. *Mathematics*, Vol. 10. MDPI. https://doi.org/10.3390/math10193554

Mardiah, M., Annisa, A., & Nidya Neyman, S. (n.d.). *Aggregate Functions in Categorical Data Skyline Search (CDSS) for Multi-keyword Document Search*.

Nurhidayah. (2024). Nurhidayah+2987. *Jurnal Sains Student Research*, (6).

Pham, B., Jovanovic, J., Bagheri, E., Antony, J., Ashoor, H., Nguyen, T. T., … Tricco, A. C. (2021). *Text mining* to support abstract screening for knowledge syntheses: a semi-automated workflow. *Systematic Reviews*, *10*(1). https://doi.org/10.1186/s13643-021-01700-x

Sabna, E. (2020). Penerapan *Text Mining* Untuk Pengelompokan Penelitian Dosen. *Jurnal Ilmu Komputer*, *9*(2), 161–164. https://doi.org/10.33060/jik/2020/vol9.iss2.183

Uray Nur Khadijah, & Nuri Cahyono. (2024). Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan *Latent dirichlet allocation* (LDA). *The Indonesian Journal of Computer Science*, *13*(4). https://doi.org/10.33022/ijcs.v13i4.3816