



Pemodelan Topik dengan *Latent-Dirichlet Allocation* untuk penemuan *State-of-the-Art* dalam Keilmuan

Timothy Dillan Tandjung

21917022

Tesis diajukan sebagai syarat untuk meraih gelar Magister Komputer

Konsentrasi Sains Data

Program Studi Informatika Program Magister

Fakultas Teknologi Industri

Universitas Islam Indonesia

2023

Lembar Pengesahan Pembimbing

Pemodelan Topik dengan *Latent-Dirichlet Allocation* untuk penemuan *State-of-the-Art*
dalam Keilmuan

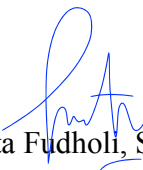
Timothy Dillan Tandjung

2023



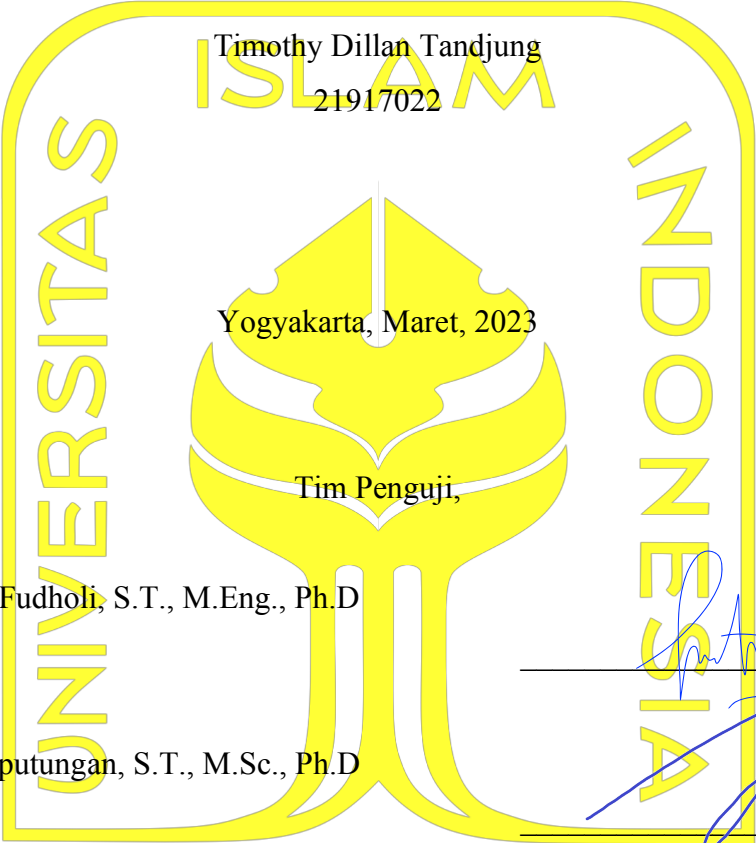
Yogyakarta, Januari, 2023

Pembimbing


Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D

Lembar Pengesahan Penguji

Pemodelan Topik dengan *Latent-Dirichlet Allocation* untuk penemuan *State-of-the-Art*
dalam Keilmuan

		
	Timothy Dillan Tandjung 21917022	
	Yogyakarta, Maret, 2023	
	Tim Penguji,	
Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D Ketua		
Irving Vitra Paputungan, S.T., M.Sc., Ph.D Anggota I		
Chandra Kusuma Dewa, S.Kom., M.Cs., Ph.D Anggota II		

Mengetahui,

Ketua Program Studi Informatika Program Magister
Universitas Islam Indonesia


Irving Vitra Paputungan, S.T., M.Sc., Ph.D.

Abstrak

Pemodelan Topik dengan *Latent-Dirichlet Allocation* untuk penemuan *State-of-the-Art* dalam Keilmuan

Mengembangkan ilmu pengetahuan adalah ambisi dalam dunia akademis. Dalam melakukan hal tersebut, penting untuk menemukan *state-of-the-art* dalam keilmuan sebagai dasar untuk pengembangan ilmu. Pendekatan penemuan *state-of-the-art* yang ada saat ini cenderung *language-specific* dan belum dapat memfasilitasi penemuan *state-of-the-art* secara otomatis dan intuitif. Tujuan penelitian ini adalah untuk menghasilkan sebuah perangkat lunak yang dapat mempermudah proses penemuan *state-of-the-art* pada topik apapun dalam dunia akademis dan yang juga *language-agnostic*. Sebuah perangkat lunak yang dapat melakukan pengumpulan artikel dari berbagai basis data untuk topik apapun secara otomatis dibangun. Konten artikel yang dikumpulkan lalu dipraproses dengan *tokenization*, *token cleansing*, penghapusan *stopwords*, dan *lemmatization*. Konten yang telah dipraproses, lalu dikonversi ke sebuah matriks dokumen-frasa numerik dan dianalisis dengan model *latent-dirichlet allocation* (LDA) dan *bidirectional encoder representations from transformers* (BERT) untuk melakukan penemuan dan pelabelan topik secara otomatis. Sebuah metode penyaringan luaran topik berbasis *entity linking* juga diusulkan dalam penelitian ini dimana luaran topik dari model difiltrasi menggunakan basis data pengetahuan agar memastikan luaran topik sesuai. Luaran topik lalu divisualisasi dengan *nested bubble* dan *line chart* untuk memvisualisasikan topik terkini, kesenjangan, dan tren evolusi dalam penelitian. Sebuah survei yang disebarakan untuk menilai aspek antarmuka dan relevansi topik dari luaran penelitian menunjukkan bahwa antarmuka perangkat lunak yang telah dikembangkan mudah untuk dinavigasi dan informasi yang ditunjukkan mudah untuk dicerna. Selain itu, hasil survei juga mengindikasikan bahwa topik-topik yang dihasilkan relevan dan sesuai dengan konten artikel yang diproses, serta visualisasi yang dihasilkan membantu dalam memahami *state-of-the-art* dan peta penelitian pada topik yang diberikan. Hasil survei ini menunjukkan bahwa luaran penelitian adekuat untuk digunakan dalam analisis preliminar penelitian dengan presisi dan relevansi yang tinggi.

Kata kunci

pemodelan topik, penemuan *state-of-the-art*, pengembangan ilmu pengetahuan; *latent-dirichlet allocation*

Abstract

Topic Modeling with Latent-Dirichlet Allocation for the Discovery of the State-of-the-art in Research

The advancement of knowledge is an ambition in the academic realm. In pursuing such, it is crucial to uncover the state-of-the-art in a given field as a foundation for furthering knowledge. Currently, the existing state-of-the-art discovery approach is language-specific and lacks the ability to facilitate state-of-the-art discovery in an automatic and intuitive manner. The aim of this study is to produce software that simplifies the process of discovering the state-of-the-art on any academic topic, while being language-agnostic. The software automatically collects articles from various databases for any given topic. The collected article content is then preprocessed through tokenization, token cleansing, stopword removal, and lemmatization. The preprocessed content is then converted into a numeric document-phrase matrix and analyzed using latent-dirichlet allocation (LDA) and bidirectional encoder representations from transformers (BERT) to automatically discover and label topics. A novel topic filtering method based on entity linking is also included in the study, where the topic output from the model is filtered using a knowledge database to ensure the output topics are relevant. The topic output is visualized through nested bubble and line charts to visualize current topics, gaps, and trends in research evolution. A survey distributed to assess the interface and topic relevance of the research output indicates that the developed software interface is user-friendly and easy to navigate, and the information presented is easily comprehensible. Additionally, the survey results also indicate that the produced topics are relevant and consistent with the processed article content, and the generated visualization aids in understanding the state-of-the-art and research map for the given topic. The survey results demonstrate that the research output is adequate for use in preliminary research analysis with high precision and relevance.

Keywords

topic modeling, state-of-the-art discovery, development of knowledge; latent-dirichlet allocation

Pernyataan Keaslian Tulisan

Dengan ini saya menyatakan bahwa tesis ini merupakan tulisan asli dari penulis, dan tidak berisi material yang telah diterbitkan sebelumnya atau tulisan dari penulis lain terkecuali referensi atas material tersebut telah disebutkan dalam tesis. Apabila ada kontribusi dari penulis lain dalam tesis ini, maka penulis lain tersebut secara eksplisit telah disebutkan dalam tesis ini.

Dengan ini saya juga menyatakan bahwa segala kontribusi dari pihak lain terhadap tesis ini, termasuk bantuan analisis statistik, desain survei, analisis data, prosedur teknis yang bersifat signifikan, dan segala bentuk aktivitas penelitian yang dipergunakan atau dilaporkan dalam tesis ini telah secara eksplisit disebutkan dalam tesis ini.

Segala bentuk hak cipta yang terdapat dalam material dokumen tesis ini berada dalam kepemilikan pemilik hak cipta masing-masing. Apabila dibutuhkan, penulis juga telah mendapatkan izin dari pemilik hak cipta untuk menggunakan ulang materialnya dalam tesis ini.

Yogyakarta, Januari, 2023

Timothy Dillan Tandjung, BSc (Hons)

Daftar Publikasi

Publikasi yang menjadi bagian dari tesis

Publikasi berikut menjadi bagian dari Bab 2.

Tandjung, D. T., & Fudholi, H. D. (2022). Topic Modeling with Latent-Dirichlet Allocation for the Discovery of State-of-the-Art in Research: A Literature Review. *Journal of Harbin Institute of Technology*, 54(8), 335–341. <https://doi.org/10.11720/JHIT.54082022.32>

Kontributor	Jenis Kontribusi
Timothy Dillan Tandjung	Memberi ide (70%) Menulis <i>paper</i> (80%)
Dhomas Hatta Fudholi	Memberi ide dan saran (30%) Mengulas <i>paper</i> (20%)

Publikasi berikut menjadi bagian dari Bab 3, dimana metodologi yang diterapkan pada publikasi digunakan juga dalam penelitian ini.

Dillan, T., & Fudholi, D. H. (2022). What can we learn from MOOC: A sentiment analysis, n-gram, and topic modeling approach. *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, 1–8. <https://doi.org/10.1109/ICITDA55840.2022.9971305>

Kontributor	Jenis Kontribusi
Timothy Dillan Tandjung	Memberi ide (70%) Mendesain eksperimen (80%) Menulis <i>paper</i> (80%)
Dhomas Hatta Fudholi	Memberi ide dan saran (30%) Mendesain eksperimen (20%) Mengulas <i>paper</i> (20%)

Halaman Kontribusi

Dalam proses penelitian dan penulisan tesis ini, penulis mengucapkan terima kasih kepada Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D yang telah memberikan bimbingan dan masukan terhadap ide dan pelaksanaan penelitian ini secara keseluruhan. Kontribusi Bapak Fudholi merupakan kontribusi signifikan yang berpengaruh dalam tesis ini. Penulis sangat berterima kasih atas bantuan dan dukungan yang telah diberikan oleh Bapak Fudholi selama proses penelitian dan penulisan laporan tesis ini.

Halaman Persembahan

Pada halaman persembahan ini, penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada orang tua penulis yang telah memberikan dukungan selama studi di Program Studi Informatika Program Magister. Penulis juga mengucapkan terima kasih kepada kakak penulis yang telah memberikan dukungan dan masukan yang sangat bermanfaat. Penulis juga ingin menyampaikan terima kasih kepada Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D, selaku dosen pemimbing yang telah memberikan bimbingan, dukungan, dan segala masukan yang sangat berguna dalam penyelesaian penelitian ini. Kontribusi Bapak Fudholi merupakan bagian yang tidak terpisahkan dari keberhasilan penelitian ini. Penulis mengucapkan terima kasih kepada pihak yang telah memberikan bantuan finansial berupa dana hibah penelitian tesis magister yang diberikan oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi selama proses studi ini. Penulis juga ingin menyampaikan terima kasih kepada Ibu Izzati Muhimmah, S.T., M.Sc., Ph.D, selaku ketua Program Studi Informatika Program Magister Fakultas Teknologi Industri Universitas Islam Indonesia. Penulis menghargai kepercayaan Ibu Izzati Muhimmah yang telah menerima penulis sebagai murid di Universitas Islam Indonesia dan memberikan kesempatan kepada penulis untuk melakukan studi S2. Penulis juga mengucapkan terima kasih atas segala bantuan dan dukungan yang telah diberikan oleh Ibu Muhimmah selama proses studi di Universitas Islam Indonesia. Sebagai penutup, penulis mengucapkan terima kasih kepada semua pihak yang telah membantu dan memberikan dukungan selama proses studi ini. Penulis menghargai setiap perhatian dan bantuan yang telah diberikan.

Kata Pengantar

Assalamu'alaikum warahmatullahi wabarakatuh.

Puji syukur penulis panjatkan ke hadirat Allah Subhanahu wa ta'ala yang telah memberikan segala rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul "Pemodelan Topik dengan *Latent-Dirichlet Allocation* untuk Penemuan *State-of-the-Art* dalam Keilmuan" ini. Laporan tesis ini merupakan salah satu syarat untuk menyelesaikan program Magister Strata 2 (S2) di Program Studi Informatika Fakultas Teknologi Industri Universitas Islam Indonesia. Dalam penyusunan laporan tesis ini, penulis telah banyak menerima bantuan, saran, dan dukungan dari berbagai pihak yang tidak dapat kami sebutkan satu persatu. Namun, pada kesempatan ini, penulis ingin menyampaikan terima kasih kepada:

1. **Bapak Dthomas Hatta Fudholi, S.T., M.Eng., Ph.D**, selaku dosen pembimbing yang telah memberikan bimbingan, masukan, dan dukungan selama proses penyusunan tesis ini.
2. **Bapak Fathul Wahid, S.T., M.Sc., Ph.D**, sebagai Rektor Universitas Islam Indonesia, yang telah memberikan kepercayaan kepada penulis selama proses studi.
3. **Bapak Irving Vitra Paputungan, S.T., M.Sc., Ph.D**, sebagai ketua Program Studi Informatika Program Magister Fakultas Teknologi Industri Universitas Islam Indonesia, yang telah memberikan dukungan kepada penulis selama proses studi.
4. **Ibu Izzati Muhimmah, S.T., M.Sc., Ph.D**, selaku ketua Program Studi Informatika Program Magister Fakultas Teknologi Industri Universitas Islam Indonesia pada saat penulis pertama kali menjadi murid magister, atas kebaikan hatinya untuk memberikan kesempatan kepada penulis untuk mengikuti program S2 di universitas ini.
5. Seluruh dosen penulis di Program Studi Informatika Program Magister Fakultas Teknologi Industri Universitas Islam Indonesia yang telah memberikan pengalaman yang menyenangkan selama belajar di UII: **Ahmad Raf'ie Pratama, S.T., M.I.T., Ph.D., Dr. Syarif Hidayat, S.Kom., M.I.T., Chandra Kusuma Dewa, S.Kom., M.Cs., Ph.D., Lizda Iswari, S.T., M.Sc., Mukhammad Andri Setiawan, S.T., M.Sc., Ph.D., Dr. H. Nur Kholis, S.Ag., SEI., M.Sh.Ec., Hendra Setiawan, S.T., M.T., Ph.D., Prof. Dr. Jaka Nugraha, S.Si., M.Si, Prof. Dr. apt. Yandi Syukri, S.Si., M.Si., Dr.techn Rohmatul Fajriyah, S.Si, M.Si., Zainudin Zukhri, S.T.,**

M.IT., Dr. Ahmad Luthfi, S.Kom., M.Kom., Om Jerri Igro, S.E., M.M., Mbak Dra. Asti Maharti Niken Sari.

6. Pimpinan dan rekan kerja di PT. Tokopedia yang telah memberikan keleluasaan waktu untuk mengerjakan penelitian ini: **Bapak William Tanuwijaya, Bapak Yoas Yohanda,** dan **Bapak Mulyadi Setiawan.**
7. Keluarga yang memberikan pemantik diskusi yang selalu hangat dan produktif untuk menghasilkan penelitian yang lebih baik: Ayah, Bunda, dan kakak.

Laporan tesis ini terbatas. Karenanya, penulis yakin masih terdapat kekurangan yang terdapat dalam laporan tesis ini. Untuk itu, penulis sangat mengharapkan saran dan kritik dari pembaca guna dimungkinkannya penemuan penelitian yang lebih baik bagi kemajuan ilmu pengetahuan di bidang ini.

Wassalamu'alaikum warahmatullahi wabarakatuh.

Daftar Isi

Lembar Pengesahan Pembimbing	i
Lembar Pengesahan Penguji.....	ii
Abstrak.....	iii
Abstract.....	iv
Pernyataan Keaslian Tulisan	v
Daftar Publikasi	vi
Halaman Kontribusi.....	vii
Halaman Persembahan	viii
Kata Pengantar.....	ix
Daftar Isi	xi
Daftar Tabel.....	xiv
Daftar Gambar	xv
Glosarium	xvii
BAB 1 Pendahuluan	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Tujuan Penelitian.....	4
1.4 Kontribusi Penelitian.....	4
1.5 Batasan Penelitian	5
1.6 Struktur Laporan	6
BAB 2 Tinjauan Pustaka	7
2.1 Penelitian	7
2.2 State-of-the-Art	9
2.3 Tinjauan Pustaka Sistematis.....	9
2.4 Analisis Bibiliometrik	12

2.5	Pemodelan Topik.....	12
BAB 3 Metodologi		22
3.1	Langkah Penelitian	22
3.2	Tinjauan Pustaka	22
3.3	Penentuan Algoritma Pemodelan Topik	23
3.3.1	Pengumpulan Data	24
3.3.2	Prapemrosesan Data	24
3.3.3	Vektorisasi.....	25
3.3.4	Pemodelan	26
3.3.5	Pelabelan Topik.....	26
3.3.6	Filtrasi Frasa Kandidat	27
3.3.7	Perbandingan Model LDA dan LSA.....	28
3.4	Implementasi Perangkat Lunak	28
3.5	Pengumpulan Data	28
3.5.1	CORE	31
3.5.2	arXiv.....	32
3.5.3	Emerald	32
3.5.4	ScienceOpen.....	33
3.5.5	Garuda	33
3.6	Pemodelan pada Runtime	34
3.6.1	Prapemrosesan Data	34
3.6.2	Vektorisasi.....	34
3.6.3	Pemodelan	35
3.6.4	Pelabelan Topik.....	35
3.6.5	Filtrasi Frasa Kandidat	36
3.7	Visualisasi	36
3.8	Evaluasi Perangkat Lunak	37

3.9	<i>Deployment</i> Perangkat Lunak	39
BAB 4 Hasil dan Pembahasan		41
4.1	Penentuan Algoritma	41
4.1.1	LDA	41
4.1.2	LSA	44
4.1.3	Perbandingan Model LDA dan LSA	45
4.2	Hasil Implementasi Perangkat Lunak	47
4.2.1	Pencarian Standar	47
4.2.2	Pencarian <i>Advanced</i>	48
4.2.3	Visualisasi <i>Nested Bubble Chart</i>	49
4.2.4	Visualisasi <i>Line Chart</i>	51
4.2.5	Riwayat Pencarian	52
4.2.6	Detil Pencarian dan Pemodelan	53
4.3	Hasil Evaluasi Perangkat Lunak	55
4.3.1	Hasil Pengujian Awal	55
4.3.2	Hasil Survei Perangkat Lunak	61
BAB 5 Kesimpulan dan Saran		79
5.1	Kesimpulan	79
5.2	Saran	80
Daftar Pustaka		81
LAMPIRAN A		90

Daftar Tabel

Tabel 2.1 Ringkasan dan Ulasan Kritis Pada Tinjauan Pustaka.....	17
Tabel 3.1 Butir Kuesioner Evaluasi Perangkat Lunak.	38
Tabel 4.1 Frasa Representatif untuk Setiap Topik LDA sebelum dan sesudah Filtrasi dan Normalisasi.	43
Tabel 4.2 Frasa Representatif untuk Setiap Topik LSA sebelum dan sesudah Filtrasi dan Normalisasi.	45
Tabel 4.3 Perbandingan Hasil antara Model LDA dan LSA.	46
Tabel 4.4 Sampel Konten Dokumen CSV dari Perangkat Lunak.	54
Tabel 4.5 Luaran Topik Perangkat Lunak untuk Bidang Pengenalan Gambar.	58
Tabel 4.6 Sampel Konten Dokumen CSV untuk Kueri Pencarian <i>Circular Economy</i>	67
Tabel 4.7 Luaran Topik Perangkat Lunak untuk Bidang <i>Circular Economy</i>	69

Daftar Gambar

Gambar 2.1 Diagram metode ilmiah (Dodig Crnkovic, 2002).....	7
Gambar 2.2 Tinjauan pustaka sistematis (Wahono, 2015).....	11
Gambar 2.3 Notasi plat LDA (Mathur et al., 2016).	14
Gambar 3.1 Metode penelitian.	22
Gambar 3.2 Pencarian artikel pada basis data ScienceDirect tanpa akun atau akses <i>proxy</i> .29	
Gambar 3.3 Pembukaan abstrak dan permintaan HTTP berulang pada basis data ScienceDirect.	29
Gambar 3.4 Pencarian artikel pada basis data JSTOR.	30
Gambar 3.5 Pencarian artikel pada basis data Google Scholar.	31
Gambar 3.6 Arsitektur perangkat lunak.	40
Gambar 4.1 Nilai koherensi model LDA untuk berbagai nilai K.....	41
Gambar 4.2 Peta intertopik distance model LDA.....	42
Gambar 4.3 Frasa dari setiap topik yang dihasilkan oleh model LDA.....	42
Gambar 4.4 Nilai koherensi model LSA untuk berbagai nilai K.	44
Gambar 4.5 Frasa dari setiap topik yang dihasilkan oleh model LSA.	44
Gambar 4.6 Fitur pencarian standar.	47
Gambar 4.7 Fitur pencarian <i>advanced</i> (a).	48
Gambar 4.8 Fitur pencarian <i>advanced</i> (b).	49
Gambar 4.9 Sampel luaran visualisasi <i>nested bubble chart</i>	50
Gambar 4.10 Sampel luaran rincian visualisasi <i>nested bubble chart</i>	51
Gambar 4.11 Sampel luaran visualisasi <i>line chart</i>	52
Gambar 4.12 Sampel riwayat pencarian.	53
Gambar 4.13 Sampel detail pencarian dan pemodelan.	54
Gambar 4.14 Parameter yang digunakan untuk melakukan pencarian pada perangkat lunak.	55
Gambar 4.15 Tangkapan layar detail parameter pemodelan dari perangkat lunak.	56
Gambar 4.16 Visualisasi <i>nested bubble chart</i> dari hasil pemodelan topik dalam bidang pengenalan gambar.	57
Gambar 4.17 Visualisasi <i>line chart</i> dari hasil pemodelan topik dalam bidang pengenalan gambar.	60
Gambar 4.18 Distribusi fakultas responden dalam survei evaluasi luaran penelitian.	61

Gambar 4.19 Distribusi nilai kemudahan navigasi pada aplikasi dalam survei evaluasi luaran penelitian.	62
Gambar 4.20 Distribusi nilai kemudahan perolehan informasi pada aplikasi dalam survei evaluasi luaran penelitian.	62
Gambar 4.21 Distribusi nilai kemudahan pencernaan informasi pada aplikasi dalam survei evaluasi luaran penelitian.	63
Gambar 4.22 Distribusi nilai relevansi topik pada aplikasi dalam survei evaluasi luaran penelitian.	64
Gambar 4.23 Distribusi nilai keterkaitan artikel dengan label topik pada aplikasi dalam survei evaluasi luaran penelitian.....	64
Gambar 4.24 Distribusi nilai kegunaan luaran dan visualisasi pada aplikasi dalam survei evaluasi luaran penelitian.	65
Gambar 4.25 Tangkapan layar detil parameter pemodelan dari perangkat lunak.	67
Gambar 4.26 Visualisasi <i>nested bubble chart</i> dari hasil pemodelan topik dalam bidang <i>circular economy</i>	73
Gambar 4.27 Visualisasi <i>line chart</i> dari hasil pemodelan topik dalam bidang <i>circular economy</i>	75
Gambar 4.28 Visualisasi peta penelitian dari <i>Open Knowledge Maps</i> dalam bidang <i>circular economy</i> (Open Knowledge Maps, 2023).	76

Glosarium

BERT	- Bidirectional Encoder Representations from Transformers
LDA	- Latent Dirichlet Allocation
LSA	- Latent Semantic Analysis
NLP	- Natural Language Processing

BAB 1

Pendahuluan

1.1 Latar Belakang

Karl Popper, seorang filsuf sains dan akademisi yang paling berpengaruh pada abad ke-20 (Thornton, 2022), menyadari bahwa semua ilmu pengetahuan harus berkembang. Teori perlu dibuat semapan mungkin agar dapat menjelaskan realitas. Untuk mengembangkan teori, budaya penemuan “*state-of-the-art research*” yang menjadi inti dari setiap penelitian merupakan salah satu hal yang paling penting untuk dilakukan (Carlisle et al., 1999; Popper, 1959; Revere et al., 2003). Dengan mengetahui perkembangan sebuah topik penelitian atau bidang ilmu tertentu, celah dan ruang untuk perbaikan dan pengembangan lebih lanjut dapat dilakukan untuk memajukan ilmu pengetahuan.

Upaya untuk menemukan *state-of-the-art* harus dan telah menjadi budaya akademik yang terus dilakukan oleh para akademisi (Carlisle et al., 1999; Kamogelo & Lucky, 2022; Revere et al., 2003; Southern Denmark University, 2021). Pada era digital dimana informasi dan media komunikasi ilmiah terbuka dan dapat diakses secara mudah, maka hal ini seharusnya dapat dilakukan dengan mudah. Namun, yang menjadi masalah adalah ketiadaannya wawasan yang dapat diperoleh seorang peneliti. Meski dengan banyaknya *platform* yang menyediakan jutaan riset secara gratis dan mudah, semua *platform* tersebut belum mampu menjawab kebutuhan pokok seorang peneliti, yaitu untuk mengetahui teori apa yang paling baru saat ini, dan apa perkembangan terkini dari teori yang ada. Pengembangan teori menjadi sulit karena kesusahan yang dihadapi (Ameen et al., 2019; Breitinger et al., 2019; Carlisle et al., 1999; Chen et al., 2016; Jasimuddin et al., 2005; Kamogelo & Lucky, 2022; Revere et al., 2003).

Untuk mengatasi masalah ini, terdapat beberapa upaya yang telah dilakukan oleh peneliti-peneliti untuk menggunakan teknologi dalam pencarian penelitian *state-of-the-art*. Salah satunya adalah metode yang dikemukakan oleh Oliveira et al., (2019). Di riset tersebut, mereka menggunakan analisis bibliometrik untuk memetakan riset *state-of-the-art* dan tren pada literatur. Metode yang ditawarkan membutuhkan peneliti untuk melakukan tahapan-tahapan tinjauan pustaka sistematis, seperti penentuan basis data pencarian, pencarian artikel dengan *search string* yang telah ditentukan, dan lalu mengekstraksi setiap data bibliometris (judul artikel, tahun artikel dibuat, nama jurnal, nama penulis, dan lainnya) untuk lalu dimasukkan dalam sebuah perangkat lunak seperti VOSViewer. Perangkat lunak tersebut lalu

dapat membantu peneliti dalam melakukan analisis bibliometris untuk memberikan wawasan terhadap tren dan hubungan antara penelitian dari data yang diekstraksi (Oliveira et al., 2019).

Selain penggunaan perangkat lunak tersebut, adapun pendekatan lainnya yang menggunakan algoritma pemodelan topik untuk mengekstraksi topik dominan dari topik yang dicari. Pemodelan topik sendiri merupakan sebuah cara untuk mengidentifikasi topik-topik laten dalam dokumen, dimana setiap topik yang dihasilkan direpresentasikan sebagai distribusi probabilitas atas kosakata dokumen. Dalam implementasinya, pemodelan topik telah digunakan untuk berbagai hal. Peringkasan pendapat orang terhadap sebuah peristiwa (semisal reaksi terhadap peluncuran produk baru atau berita kontroversial), identifikasi topik yang paling banyak menerima perhatian dalam domain tertentu, dan untuk mengidentifikasi evolusi topik dari teks yang ada (Smatana & Butka, 2019). Kemampuan pemodelan topik yang mampu mengatur, memahami, dan meringkas kumpulan informasi tekstual dari berbagai koleksi teks ke dalam bentuk pola topikal yang terstruktur, membuatnya sangat baik dalam melakukan analisis dan menemukan wawasan pada teks.

Mengetahui implementasi dan kemampuan pemodelan topik, beberapa penelitian telah menggunakan pemodelan topik untuk menemukan *state-of-the-art*. Pemodelan topik, terutama LDA (*Latent-dirichlet Allocation*), sebuah algoritma pemodelan topik yang *unsupervised* dan *language-agnostic* telah digunakan untuk mengekstraksi topik-topik dominan yang paling banyak menerima perhatian dan mengidentifikasi tren topik dalam topik penelitian tertentu. Kemampuan LDA yang *unsupervised* dan tidak terbatas pada bahasa tertentu membuatnya sebuah pendekatan yang memiliki kemampuan dan potensi besar untuk menganalisis konten artikel riset. Luaran pemodelan topik tersebut dapat merepresentasikan topik riset terbaru dalam penelitian, dan juga dapat menunjukkan evolusi topik yang diteliti dalam penelitian tertentu.

Namun dengan pendekatan-pendekatan ini, proses yang dilakukan masih manual dan membutuhkan keahlian dalam mepraproses data yang ingin digunakan untuk analisis. Peneliti perlu melakukan pencarian data, ekstraksi data, dan integrasi data ke dalam sebuah perangkat lunak secara mandiri (Oliveira et al., 2019). Pendekatan-pendekatan ini belum memberikan sebuah antarmuka pencarian riset atau topik terkini dari topik yang ditentukan dari berbagai basis data ilmiah dan memberikan wawasan serta informasi tentang pengembangan terkini secara otomatis.

Kemampuan dan potensi besar dari pemodelan topik membuka banyak peluang baru untuk mengembangkan implementasi pemodelan topik, apalagi dengan keperluan penemuan

state-of-the-art dengan cepat dan tepat dari volume data yang terus bertumbuh secara eksponensial. Namun dengan pendekatan sekarang, proses yang dilakukan masih manual dan membutuhkan keahlian dalam memproses data yang ingin digunakan untuk analisis (Oliveira et al., 2019). Peneliti perlu melakukan pencarian data, ekstraksi data, dan integrasi data ke dalam sebuah perangkat lunak secara mandiri. Terkadang, peneliti juga hanya akan mencari data dari satu sumber saja, yang dapat membatasi informasi mengenai sebuah topik. Pendekatan-pendekatan saat ini juga belum mempunyai sebuah antarmuka pencarian topik terkini dari berbagai basis data ilmiah dan memberikan wawasan serta informasi tentang pengembangan terkini secara otomatis. Pendekatan analisis bibliometris saat ini juga masih hanya menganalisis kata kunci dari artikel yang dikumpulkan, yang dapat menghasilkan luaran analisis yang salah jika kata yang digunakan tidak merefleksikan konten artikel dengan baik. Selain itu, pendekatan saat ini belum mampu memberikan gambaran tentang tren evolusi topik dalam riset, informasi penting yang dapat mengungkapkan perubahan paradigma penelitian dari waktu-ke-waktu.

Mempertimbangkan penemuan *state-of-the-art* yang masih menjadi kebutuhan yang penting dalam dunia akademis, serta keterbatasan dari pendekatan-pendekatan sebelumnya, penelitian ini bertujuan untuk mengembangkan pendekatan baru dalam penemuan *state-of-the-art*. Dikarenakan kemampuannya yang dapat menganalisis kumpulan konten artikel secara *unsupervised* dan ekstraksi topik-topik laten dari kumpulan artikel, serta pemberian informasi lebih lanjut seperti evolusi topik dari waktu-ke-waktu, penelitian ini mengusulkan sebuah kerangka penemuan *state-of-the-art* yang mengembangkan pendekatan terdahulu berbasis LDA.

Penelitian ini menghasilkan sebuah perangkat lunak yang dapat melakukan pengumpulan artikel dari berbagai basis data akademis menggunakan metode-metode information retrieval (IR). Konten artikel yang dikumpulkan lalu dipraproses, yang meliputi proses *tokenization*, *token cleansing*, penghapusan *stopwords*, dan *lemmatization* (Dhingra, 2021). Sebuah model LDA lalu akan dibangun secara dinamis berdasarkan konten artikel yang dikumpulkan, untuk mendapatkan topik-topik laten dalam penelitian, dan luaran model LDA diperkaya dengan penggunaan model *bidirectional encoder representations from transformers* (BERT) dan penyaringan topik dengan metode *entity linking*, untuk memastikan bahwa luaran topik intuitif dan merepresentasikan topik-topik yang nyata. Luaran topik lalu divisualisasi dengan sebuah *nested bubble chart* dan *line chart* agar dapat menunjukkan topik *state-of-the-art*, kesenjangan, dan tren evolusi dalam penelitian, yang dapat membantu peneliti dalam proses analisis preliminar penelitian.

1.2 Rumusan Masalah

Penemuan *state-of-the-art* merupakan tujuan utama dari dunia akademisi yang memiliki banyak kesulitan. Masalah yang dihadapi dalam proses penemuan *state-of-the-art* dalam penelitian adalah kesulitan dalam mengakses dan menyaring informasi yang relevan. Pendekatan-pendekatan seperti metode tinjauan pustaka sistematis, analisis bibliometrik, dan pemodelan topik yang telah digunakan sebelumnya memerlukan peneliti untuk melakukan tahapan-tahapan tinjauan pustaka sistematis dan mengekstraksi data bibliometrik secara manual, sehingga memerlukan waktu yang cukup lama dan kerja yang berulang. Pendekatan semi-otomatis seperti analisis bibliometrik juga hanya dapat menganalisis kumpulan kata kunci artikel, yang rentan dapat menghasilkan luaran yang kurang sesuai dengan konten artikel; dan kemampuannya hanya diuji untuk konten artikel berbahasa Inggris. Berdasarkan latar belakang yang telah diuraikan, adapun beberapa rumusan masalah yang diangkat dalam penelitian ini, diantaranya:

- a) Bagaimana cara mengimplementasikan LDA untuk mempercepat proses pencarian *state-of-the-art* dan, secara umum, analisis preliminar pada penelitian?
- b) Bagaimana cara membuat sebuah perangkat lunak yang dapat mengintegrasikan LDA dan pengambilan literatur ilmiah *open-access* untuk penemuan *state-of-the-art*?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah dihasilkannya sebuah model berbasis LDA yang terintegrasikan ke dalam sebuah antarmuka untuk memudahkan proses analisis preliminar dalam melakukan penelitian pada tahap penemuan *state-of-the-art*.

1.4 Kontribusi Penelitian

Dalam penelitian terdahulu, penemuan *state-of-the-art* dilakukan dengan peneliti mengumpulkan artikel (umumnya terbatas pada satu basis data riset), skrining, dan pemrosesan artikel dilakukan secara manual. Lalu, data akan dianalisis menggunakan analisis bibliometrik atau dengan pendekatan pembelajaran mesin seperti LDA, untuk menemukan topik-topik dominan dalam literatur. Kontribusi penelitian ini adalah pengusulan pendekatan baru berbasis pemodelan topik untuk menemukan *state-of-the-art* pada topik apapun secara otomatis.

Penelitian menggunakan pendekatan analisis yang serupa, dimana pemodelan topik dengan LDA digunakan untuk menemukan lanskap topik dari sebuah bidang ilmu. Tetapi, terdapat pengembangan sebuah mesin pencarian, dimana pengguna dapat langsung memasukkan topik yang ingin diteliti, dan perangkat lunak akan mengotomatisasi proses pencarian, pengumpulan, skrining, dan pemrosesan artikel untuk topik yang ditentukan. Perangkat lunak ini mampu mengumpulkan konten artikel dari berbagai basis data akademis, seperti CORE, arXiv, ScienceOpen, Emerald, dan Garuda, dan menerapkan teknik pemodelan topik menggunakan LDA yang diintegrasikan dengan ekstraksi frasa secara otomatis menggunakan pola POS dan model *bidirectional encoder representations from transformers* (BERT) untuk mengidentifikasi topik-topik dalam penelitian yang agnostik terhadap bahasa.

Sebuah metode penyaringan topik berbasis *entity linking* juga diusulkan untuk memastikan validitas luaran topik. Hasil akhir dari pemodelan topik ditampilkan dalam bentuk visualisasi *nested bubble chart* dan *line chart* untuk mempermudah pemahaman tren dan kesenjangan dalam penelitian. Dengan ini, perangkat lunak dapat menunjukkan topik-topik dominan dan tren topik dalam bidang yang ditentukan. Penelitian ini diharapkan dapat mengurangi waktu yang dihabiskan oleh peneliti dalam mencari dan menganalisis hasil penelitian terkini, serta memberikan wawasan yang lebih luas tentang pengembangan terkini dalam topik penelitian yang ditentukan. Selain itu, perangkat lunak ini diharapkan dapat memberikan informasi yang bermanfaat bagi para peneliti dalam menyoroti bidang-bidang yang membutuhkan perhatian lebih dalam agenda penelitian pada topik yang ditentukan ataupun menentukan arah dan prioritas penelitian di masa yang akan datang.

1.5 Batasan Penelitian

Di dalam penelitian ini, terdapat beberapa batasan cakupan, diantaranya:

- a) Penelitian hanya akan menguji implementasi pemodelan topik untuk penemuan *state-of-the-art* dalam bidang studi ilmu komputer dan bisnis, sehingga luaran penelitian ini hanya dapat diverifikasi penggunaannya dalam bidang studi tersebut.
- b) Data diasumsikan hanya dalam bahasa Inggris dan Indonesia, sehingga proses prapemrosesan yang diaplikasikan pada data akan menggunakan model dan data yang berbahasa Inggris dan Indonesia.
- c) Proses *scraping* data hanya akan dilakukan pada basis data akademis CORE, ScienceOpen, arXiv, Emerald dan Garuda. Selain itu, data yang dikumpulkan hanya dari penelitian dengan akses terbuka untuk publik.

- d) Algoritma pemodelan topik yang akan digunakan dan dibandingkan hanya LDA dan LSA.
- e) Evaluasi model hanya akan dilakukan dengan metrik koherensi *UMass* dan basis data yang digunakan untuk seleski model adalah basis data CORE dengan topik penelitian yang telah ditentukan.
- f) Perbandingan luaran penelitian dengan perangkat lunak serupa hanya dilakukan dengan *Open Knowledge Maps*.

1.6 Struktur Laporan

Laporan tesis ini disusun menggunakan struktur sebagai berikut:

a) Pendahuluan

Penelitian dimulai dengan pendahuluan yang menjelaskan latar belakang, permasalahan, tujuan, kontribusi dan manfaat, serta batasan penelitian dan struktur laporan penelitian.

b) Tinjauan Pustaka

Laporan lalu dilanjutkan dengan tinjauan pustaka yang menguraikan literatur-literatur yang relevan dengan penelitian ini, termasuk teori dan konsep yang diperlukan untuk memahami penelitian seperti kajian tentang *state-of-the-art*, metode penemuan *state-of-the-art*, dan pemodelan topik.

c) Metodologi

Bagian metodologi menjelaskan langkah-langkah yang dilakukan dalam pelaksanaan penelitian secara detail.

d) Hasil dan Pembahasan

Bagian ini lalu menyajikan hasil penelitian yang diperoleh dari metodologi yang telah dijelaskan serta penguraian dan pembahasannya dengan detail.

e) Kesimpulan dan Saran

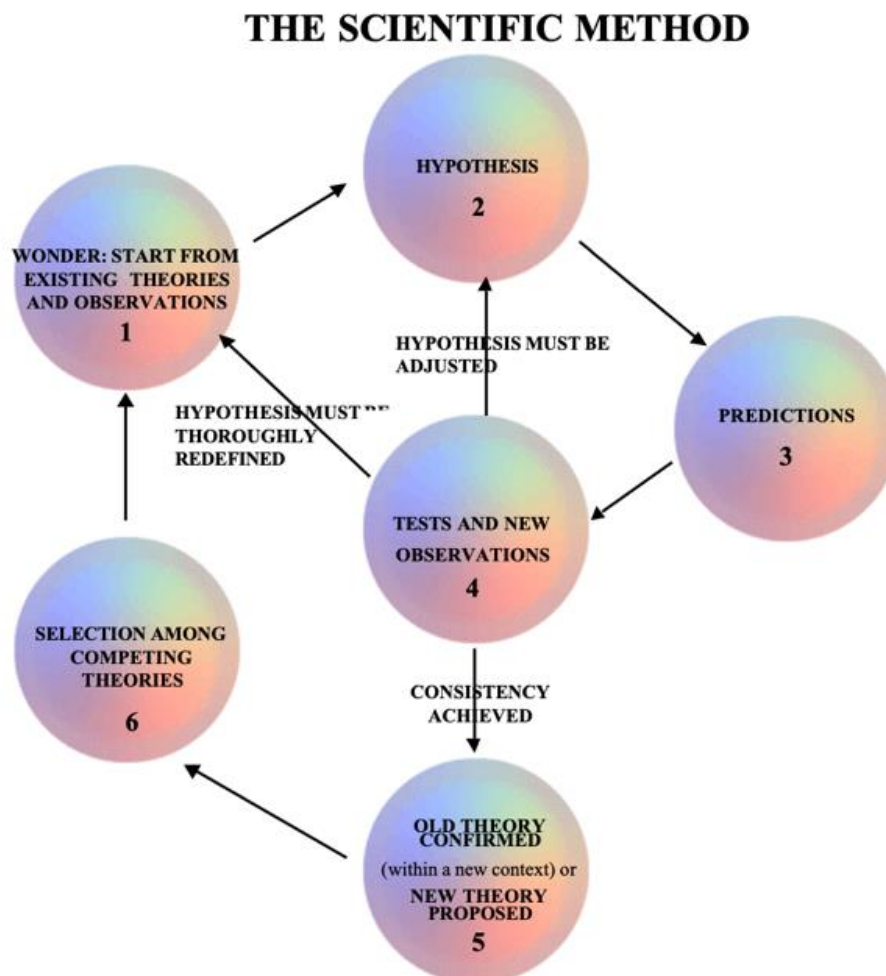
Laporan lalu ditutup dengan bagian kesimpulan dan saran yang menyajikan kesimpulan yang dapat diambil dari laporan ini, serta memberikan rekomendasi untuk penelitian selanjutnya.

BAB 2

Tinjauan Pustaka

2.1 Penelitian

Dalam melakukan penelitian, peneliti pada umumnya akan mengikuti serangkaian langkah yang diringkaskan dalam diagram alir pada Gambar 2.1. Gambar menunjukkan metode ilmiah yang digunakan oleh peneliti untuk menemukan, mengembangkan, dan atau memvalidasi teori (Dodig Crnkovic, 2002). Adapun langkah yang diikuti oleh peneliti dimulai dari tahap pertama, “*Wonder: Start from existing theories and observations*”, dimana teori dan fenomena yang ada diamati, lalu diikuti secara sekuensial hingga perumusan masalah, pembentukan hipotesis, pengujian hipotesis, hingga tahap keenam, yakni “*Selection among competing theories*” atau proses ketahanan sebuah teori diantara teori lainnya dalam menjelaskan sebuah fenomena.



Gambar 2.1 Diagram metode ilmiah (Dodig Crnkovic, 2002).

Metode ilmiah dalam gambar direpresentasikan dalam enam langkah sekuensial dan rekursif, dengan penjelasan masing-masing langkah sebagai berikut:

- 1) Pengajuan pertanyaan dalam konteks pengetahuan yang ada (teori & observasi), dimana pertanyaan tersebut bisa berupa pertanyaan yang mampu dijawab oleh teori lama, atau pertanyaan yang membutuhkan perumusan teori baru.
- 2) Formulasi hipotesis sebagai jawaban tentatif. Langkah ini melibatkan formulasi hipotesis berdasarkan teori dan pengamatan yang ada untuk menjelaskan fenomena yang diteliti dan menyediakan titik awal untuk penelitian.
- 3) Merumuskan konsekuensi dan prediksi dari hipotesis, yang nantinya digunakan untuk menguji validitas hipotesis. Prediksi dan konsekuensi ini digunakan untuk mengarahkan desain eksperimen dan pengumpulan data.
- 4) Uji hipotesis dalam eksperimen atau bidang teori tertentu. Hipotesis baru harus terbukti sesuai dengan pandangan dunia yang ada. Jika hipotesis mengarah pada kontradiksi dan menuntut perubahan radikal dalam latar belakang teoritis yang ada, hipotesis tersebut harus diuji dengan sangat hati-hati. Proses ini dapat berulang dari langkah kedua untuk memodifikasi hipotesis hingga sebuah kesepakatan diperoleh. Jika ditemukan ketidaksesuaian hipotesis maka proses harus dimulai dari awal.
- 5) Ketika konsistensi dalam hasil uji hipotesis diperoleh, hipotesis menjadi teori dan menyediakan seperangkat proposisi yang koheren yang mendefinisikan kelas fenomena baru atau konsep teoretis baru. Sebuah teori kemudian menjadi kerangka dimana pengamatan atau fakta teoritis dijelaskan dan prediksi dibuat.
- 6) Teori yang dibangun lalu bersaing dengan teori lainnya, dan melewati sebuah proses “seleksi alam”. Teori yang dapat menjelaskan fenomena yang ada dengan paling baik bertahan.

Secara keseluruhan, diagram alir pada Gambar 2.1 memberikan gambaran yang jelas dan menyeluruh tentang proses penelitian dan tahap utama yang terlibat dalam menemukan, mengembangkan, dan memvalidasi ilmu. Penelitian ini menaruh fokus pada langkah pertama dalam melakukan penelitian ilmiah, yakni pengajuan pertanyaan dalam konteks pengetahuan yang ada. Pada umumnya, perolehan konteks pengetahuan yang ada dapat dilakukan dengan memahami *state-of-the-art* dari teori yang biasanya digunakan untuk menjelaskan fenomena yang ingin diteliti.

2.2 State-of-the-Art

State-of-the-art, menurut Santos (2019), menggambarkan pengetahuan saat ini tentang fenomena yang ingin dipelajari. Penemuan *state-of-the-art* dapat memberikan gambaran yang komprehensif tentang apa yang telah dilakukan pada suatu bidang ilmu dan menunjukkan potensi untuk investigasi lebih lanjut terhadap kasus-kasus berbeda yang mungkin belum sepenuhnya dapat dijelaskan oleh teori-teori terdahulu. Merujuk ke diagram metode ilmiah, tentunya hal ini merupakan langkah awal utama yang sangat krusial dalam melakukan penelitian, karena menjadi dasar dalam merumuskan pertanyaan atau masalah penelitian (Santos, 2019).

Terdapat beberapa jenis *state-of-the-art* yang dapat diidentifikasi. Misalnya, ada *state-of-the-art* yang bersifat teoretis, yang meliputi teori dan model yang paling banyak diterima di bidang tersebut. Terdapat juga *state-of-the-art* empiris, yang mengacu pada temuan dan penemuan terbaru berdasarkan penelitian empiris. Selain itu, ada *state-of-the-art* teknologi, yang berkaitan dengan alat, teknologi, dan metodologi terbaru yang digunakan dalam bidang tersebut (Santos, 2019).

Dalam tinjauan literatur, istilah ini digunakan untuk menggambarkan kondisi pengetahuan saat ini dalam area penelitian tertentu dan untuk mengidentifikasi perkembangan terbaru. Tinjauan literatur tentang *state-of-the-art* pada umumnya meliputi gambaran kondisi pengetahuan saat ini dalam bidang atau area penelitian tertentu, dengan menyoroti perkembangan terbaru dalam penelitian. Hal ini meliputi ringkasan lengkap penelitian terbaru, metode, dan teknik yang telah diajukan dan digunakan dalam bidang tersebut, serta memberikan pandangan ke arah penelitian masa depan.

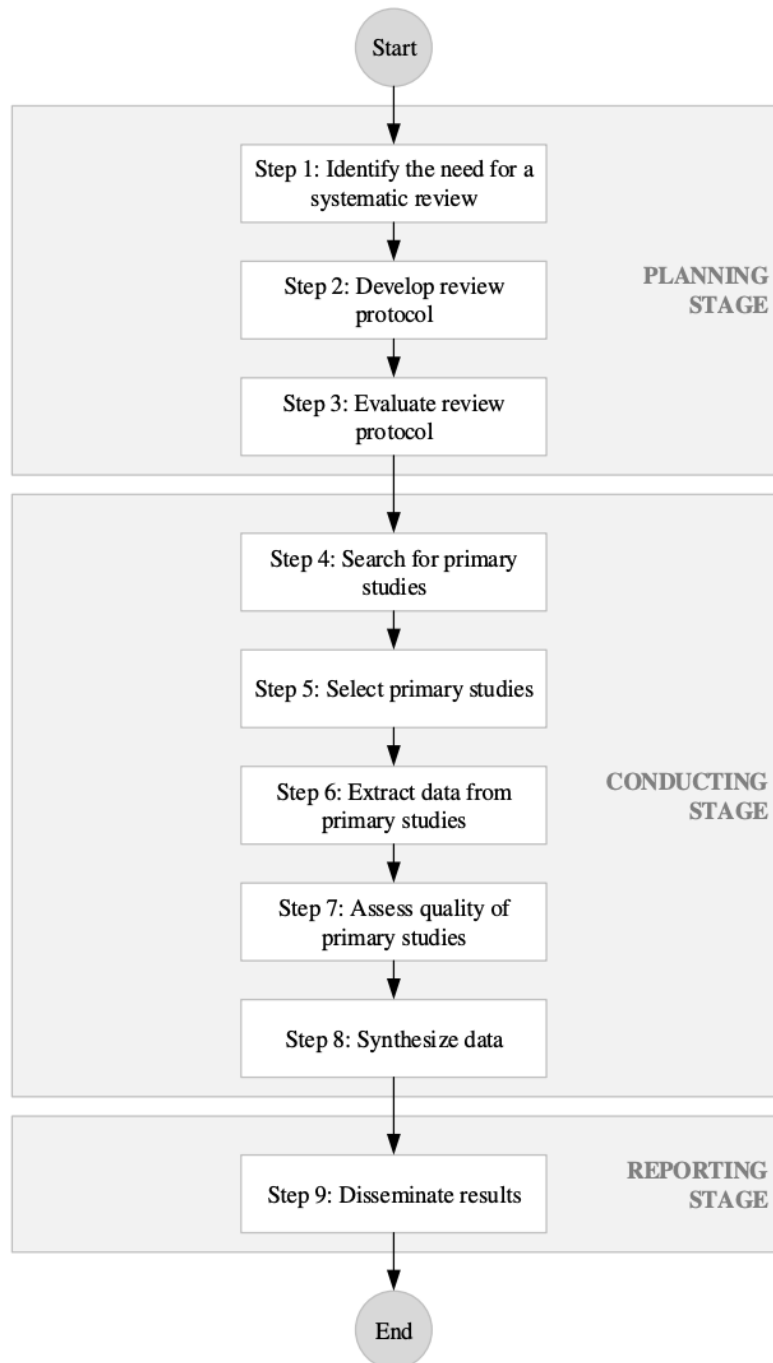
Dalam konteks penelitian ini, penelitian menaruh fokus pada penemuan *state-of-the-art* teoretis, khususnya dalam menemukan topik penelitian terbaru dalam suatu bidang tertentu. Pendekatan yang diusulkan dalam penelitian ini yang menggunakan LDA, BERT, filtrasi dengan *entity linking* dan *knowledge base*, serta pengumpulan, pemrosesan data, dan visualisasi intuitif secara otomatis, diharapkan dapat membantu proses pengungkapan topik penelitian terbaru pada bidang ilmu.

2.3 Tinjauan Pustaka Sistematis

Metode yang dikenal dan digunakan secara luas sampai saat ini untuk menemukan *state-of-the-art* adalah analisis literatur atau kerap disebut sebagai tinjauan pustaka. Terdapat berbagai definisi tinjauan pustaka, namun, tinjauan pustaka dapat didefinisikan sebagai pengumpulan, analisis, dan sintesis temuan, ide, dan perspektif dari literatur dalam topik

tertentu. Hal ini kemudian akan memberikan landasan konteks pengetahuan yang kuat untuk topik yang ingin diteliti, dan dapat membantu menunjukkan area yang belum terjangkau di mana penelitian yang diusulkan berkontribusi (Jahan et al., 2016; Levy & Ellis, 2006; Santos, 2019; Snyder, 2019). Tinjauan pustaka sendiri dapat dipecah menjadi tiga jenis, yang mencakup tinjauan sistematis, semi-sistematis, dan integratif. Dari ketiga jenis tersebut, di sebagian besar bidang ilmu pengetahuan, termasuk ilmu komputer, tinjauan pustaka sistematis telah menjadi cara yang paling berkembang dan disepakati untuk melakukan tinjauan pada literatur, dengan banyak peneliti mencatatnya sebagai standar utama tinjauan pustaka (Davis et al., 2014; Snyder, 2019).

Tinjauan pustaka sistematis merupakan tinjauan pustaka yang dilakukan secara sistematis dan *reproducible*. Tinjauan pustaka sistematis biasanya dilakukan dalam tiga tahap utama, yang terdiri dari *planning*, *conducting*, dan *reporting*, sebagaimana dijelaskan pada Gambar 2.2. Pada tahap *planning*, perlunya tinjauan pustaka, dan pengembangan protokol tinjauan (pemilihan kata kunci, basis data pencarian, dan kriteria inklusi atau eksklusi artikel). Selanjutnya pada tahap *conducting*, pengumpulan literatur dengan protokol review yang telah ditentukan sebelumnya oleh peneliti akan dilakukan. Literatur yang terkumpul akan disaring lebih lanjut untuk mengidentifikasi literatur yang relevan, setelah identifikasi literatur relevan, sintesa ide dari masing-masing literatur akan dilakukan. Lalu, tahap *reporting* adalah di mana hasil sintesis dibagikan (Snyder, 2019; Wahono, 2015).



Gambar 2.2 Tinjauan pustaka sistematis (Wahono, 2015).

Walaupun tinjauan pustaka sistematis dipuji sebagai “standar emas” untuk penemuan *state-of-the-art*, terdapat banyak tantangan dalam melakukannya. Melakukan tinjauan pustaka sistematis dapat disebut sulit dan memakan waktu, karena melibatkan proses analisis, perbandingan, evaluasi dan penghubungan antar sumber yang berbeda, yang membutuhkan berjam-jam kerja manual dalam membaca dan organisasi konten untuk mencapainya. Selain itu, karena produksi pengetahuan semakin cepat dengan kecepatan

yang luar biasa, pemilihan literatur bisa menjadi sulit ketika dihadapkan dengan jumlah kandidat yang terus bertambah banyak. Akibatnya, ini menyebabkan kesulitan dalam memahami, mengevaluasi, dan mensintesis *state-of-the-art* dari topik yang ingin diteliti (Ameen et al., 2019; Chen et al., 2016; Felizardo et al., 2012; Kamogelo & Lucky, 2022; Snyder, 2019).

2.4 Analisis Bibliometrik

Dalam upaya untuk membantu membuat proses ekstraksi dan pengertian dokumen lebih efisien dan terotomatisasi, metode untuk mendapatkan wawasan dan memahami dokumen telah menjadi riset yang diminati pada area informatika, lebih spesifiknya pada bidang scientometrik dan infometrik. Dengan kemajuan yang begitu pesat pada teknologi dan riset pada bidang ini yang terus berkembang, adapun terciptanya banyak teori dan algoritma untuk mengeksplorasi pengertian dokumen. Upaya awal untuk memahami dan memproses literatur ilmiah dimulai pada tahun 1955, di mana Garfield (1955) mengemukakan gagasan menggunakan indeks *citation* untuk mengaitkan ide di antara literatur. Saat ini, pengelompokan (*clustering*) dan kesamaan (*similarity*) dokumen dengan data bibliometrik telah menjadi salah satu cara mutakhir dalam memproses literatur ilmiah (Ballester & Penner, 2022).

Sejalan dengan visi Garfield, pengembangan perangkat lunak seperti VOSViewer atau CitNext Eplorer yang mampu memproses data bibliometrik dan melakukan analisis *co-citation*, *co-authorship*, *co-occurrence of keywords*, merupakan salah satu pendekatan yang menggunakan analisis bibliometrik untuk membantu peneliti memahami topik penelitian tertentu. Dengan pendekatan ini, peneliti perlu mengumpulkan data bibliometris dari satu atau lebih basis data ilmiah. Data lalu dapat dimasukkan ke dalam perangkat lunak, dan dengan penggunaan perangkat lunak tersebut, penemuan hubungan antar riset, pemeringkatan riset, dan analisis kata kunci dapat dilakukan dan mampu membantu dalam analisis awal pada topik riset (Oliveira et al., 2019). Namun, hasil dari analisis bibliometrik terbukti sulit untuk mewakili topik dalam *body of knowledge*, karena terbatas pada pengelompokan berdasarkan data bibliometrik (Ballester & Penner, 2022).

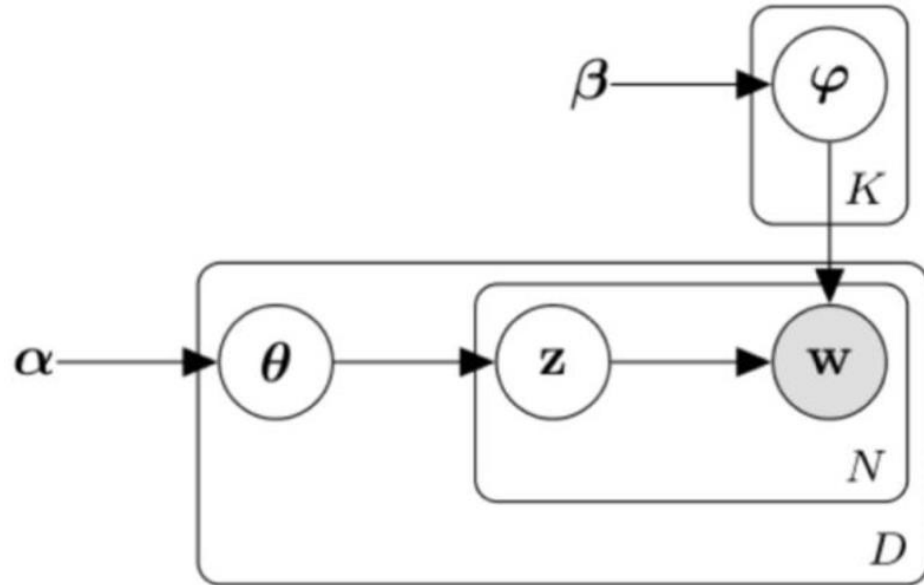
2.5 Pemodelan Topik

Sebagai alternatif, teknik pemodelan topik dari bidang ilmu komputer merupakan pendekatan lain yang dapat digunakan dalam pemahaman dokumen. Pemodelan topik adalah

metode pemodelan statistik yang dapat digunakan untuk menemukan topik-topik laten dari kumpulan dokumen (Felizardo et al., 2012; Lee et al., 2021; Nikolenko et al., 2015; Rumbut et al., 2020). Dalam bidang pemahaman literatur ilmiah, pemodelan topik telah banyak digunakan dalam mengkarakterisasi ilmu pengetahuan (Boyack et al., 2011; Gläser et al., 2017; Griffiths & Steyvers, 2004), karena kemampuannya untuk menemukan topik laten dalam kumpulan dokumen yang besar dan tidak terstruktur. Dalam pemodelan topik, topik-topik yang ada dalam dokumen diklasifikasikan berdasarkan distribusi kata-kata dalam dokumen tersebut. Dengan kata lain, topik-topik yang banyak disebutkan dalam dokumen akan memiliki bobot yang lebih tinggi dalam model pemodelan topik.

Dalam penelitian ilmiah, topik yang paling banyak dibahas dan mendapat perhatian peneliti dianggap sebagai topik yang paling penting dan signifikan dalam bidang tersebut. Pemodelan topik dapat membantu mengidentifikasi topik-topik ini secara otomatis dan objektif, tanpa memerlukan penilaian subyektif dari peneliti. Selain itu, dengan menganalisis tren dalam distribusi topik dari waktu ke waktu, pemodelan topik juga dapat membantu mengidentifikasi pergeseran dalam fokus penelitian dan kemajuan *state-of-the-art* dalam bidang ilmu.

Salah satu implementasi terkini untuk melakukan pemodelan topik menggunakan model topik probabilistik, dengan LDA dan LSA sebagai algoritma yang paling populer. Selain itu, kedua algoritma tersebut adalah algoritma yang paling umum digunakan dalam pemodelan topik dalam riset, dan telah banyak penelitian sebelumnya yang telah membandingkan kinerja keduanya dalam penemuan topik *state-of-the-art* dalam keilmuan (Ballester & Penner, 2022; Felizardo et al., 2012; Lee et al., 2021; Valdez et al., 2018). LDA pertama kali dikemukakan oleh Blei, Ng dan Jordan pada tahun 2003 dalam penelitiannya yang berjudul: “*Latent Dirichlet Allocation*”. Gambar 2.3 menunjukkan sebuah notasi plat ilustrasi proses generatif algoritma LDA. Secara konseptual, LDA memiliki asumsi dasar bahwa sebuah topik direpresentasikan oleh kata-kata yang serupa, dan bahwa setiap dokumen dapat direpresentasikan dalam beberapa topik. Selain itu, LDA juga mengasumsikan bahwa data berdistribusi dirichlet (Ballester & Penner, 2022; Lee et al., 2021). Tujuan dari LDA sendiri adalah untuk memetakan setiap dokumen ke sebuah topik yang meliputi kata-kata pada dokumen. Karena LDA memperlakukan setiap dokumen sama (sebagai *bag-of-words*), proses mengatur ulang urutan kata atau memodifikasi konten dari dokumen tidak perlu dilakukan.



Gambar 2.3 Notasi plat LDA (Mathur et al., 2016).

Secara teknis, sebuah model LDA memiliki beberapa parameter yang dipakai untuk menghasilkan topik, yang diilustrasikan pada Gambar 2.3. Di sini, α mewakili distribusi topik per-dokumen, yang mengelola distribusi topik untuk setiap dokumen dalam korpus. Nilai α yang lebih tinggi akan menyebabkan dokumen ditugaskan ke lebih banyak topik. θ adalah distribusi topik untuk dokumen, yang mewakili probabilitas dokumen saat ini ditugaskan ke topik tertentu. z menandakan topik untuk kata spesifik dalam dokumen saat ini, w menandakan kata spesifik, N adalah jumlah kata dalam suatu dokumen tertentu, D adalah jumlah dokumen, K adalah jumlah topik, ϕ adalah distribusi kata untuk setiap topik, dan terakhir, β adalah distribusi kata per-topik, yang mengelola distribusi kata untuk setiap topik. Nilai β yang lebih tinggi akan menyebabkan topik diwakili oleh lebih banyak kata. Dengan parameter-parameter diatas, cara kerja LDA dalam menghasilkan topik adalah (Ballester & Penner, 2022; Bansal, 2019; Kulshrestha, 2020; Lee et al., 2021):

Dengan jumlah dokumen D dan N jumlah kata dalam sebuah dokumen,

- a) Untuk setiap dokumen, telusuri dan tetapkan secara acak setiap kata dalam dokumen ke salah satu dari k topik (k ditentukan sebelumnya). Proses ini menginisialisasi nilai ϕ .
- b) Untuk setiap dokumen d , periksa setiap kata w dan hitung probabilitas dari:
 - a. $p(\text{topik } t \mid \text{dokumen } d)$: adalah proporsi kata dalam dokumen d yang ditugaskan ke topik t , yang juga menghasilkan pembaruan nilai pada parameter θ . Hal ini mencoba menangkap berapa banyak kata yang termasuk dalam topik t untuk

dokumen tertentu d . Jika banyak kata dari d milik t , kemungkinan besar kata w milik t . Parameter α mengatur distribusi penugasan topik ini untuk dokumen.

- b. $p(\text{kata } w \mid \text{topik } t)$: adalah proporsi penugasan ke topik t atas semua dokumen yang berasal dari kata w . Proporsi ini mencoba menangkap berapa banyak dokumen dalam topik t karena kata w . LDA mewakili dokumen sebagai campuran topik. Demikian pula, topik adalah campuran kata-kata. Jika sebuah kata memiliki probabilitas tinggi untuk menjadi topik, semua dokumen yang memiliki w akan lebih kuat terkait dengan t juga. Jika w sangat kecil kemungkinannya berada di t , dokumen yang berisi w akan memiliki probabilitas yang sangat rendah untuk berada di t , karena sisa kata dalam d akan menjadi bagian dari topik lain dan karenanya d akan memiliki nilai yang lebih tinggi. probabilitas untuk topik tersebut. Penghitungan probabilitas ini menghasilkan pembaruan pada parameter ϕ , dan distribusi ini diatur oleh hyperparameter β .
- c) Perbarui probabilitas untuk kata w milik topik t , sebagai: $p(\text{kata } w \text{ dengan topik } t) = p(\text{topik } t \mid \text{dokumen } d) * p(\text{kata } w \mid \text{topik } t)$, yang memperbarui parameter z .

Setelah probabilitas diestimasi, sekumpulan kata untuk membangun atau mewakili topik tertentu dapat dilakukan dengan memilih kata di atas ambang probabilitas tertentu atau probabilitas N teratas. Begitu pula, penugasan topik untuk setiap dokumen juga dapat dilakukan dengan cara yang serupa (Bansal, 2019; Pascual, 2019).

LSA, di sisi lain, merupakan algoritma pemodelan topik yang menggunakan proses *multistep* untuk menentukan struktur dari sebuah korpus dengan memodelkan konteks dimana sebuah kata atau terminologi muncul. Dalam pengerjaannya, LSA mengkonstruksi sebuah matriks *document-term* (DTM), yang nantinya direduksi dengan *singular value decomposition* (SVD) untuk mengaproksimasi *eigenvector* dari setiap kata dan kemiripan antar kata. Lalu, sebuah rotasi faktor diaplikasikan pada *eigenvector*, dan topik-topik akan dihasilkan dari penggerombolan *eigenvector* kata dari korpus (Valdez et al., 2018).

Dalam bidang penemuan *state-of-the-art*, beberapa penelitian telah menggunakan metode pemodelan topik LDA dan LSA untuk menemukan tren pada riset dan memberikan gambaran besar atau ringkasan terhadap sebuah topik. Semisal, dalam penelitian yang dilakukan oleh Lee et al., (2021) LDA dan analisis bibliometris digunakan untuk menemukan tren riset dalam transformasi digital di bidang manufaktur dan rekayasa. Mirip dengan penelitian tersebut, Li et al., (2019) menggunakan algoritma LDA, K-means, dan NTF-IDF (*normalized term frequency inverse document frequency*) untuk menemukan tren riset *deep learning* dalam bidang *computer vision*. LSA juga digunakan pada bidang ilmu

sosial untuk menemukan dan menganalisis pola tema dalam teks yang sangat banyak (Valdez et al., 2018). Penelitian-penelitian terdahulu ini membuktikan bahwa pemodelan topik, khususnya LDA dan LSA, dapat digunakan untuk menemukan *state-of-the-art* pada sebuah topik, dan dapat membantu peneliti untuk melakukan analisis preliminar pada penelitian.

Tetapi dari tinjauan pustaka yang telah dilakukan, belum ditemukan sebuah pendekatan yang dapat menemukan riset *state-of-the-art* pada topik penelitian apapun dengan kapabilitas memproses literatur ilmiah secara otomatis dari berbagai sumber. Sehingga, kontribusi penelitian ini adalah sebuah perangkat lunak terintegrasi dengan model topik yang mampu menemukan perkembangan terkini dari sebuah topik.

Perangkat lunak akan mengambil data bibliometrik dan konten artikel dari berbagai basis data riset, menggunakan *information retrieval* (IR), sebuah proses pengumpulan dan penyusunan dari sebuah basis data (Roshdi & Roohparvar, 2015). Lebih spesifiknya, perangkat lunak akan menggunakan teknik *web scraping*, yang merupakan cara untuk mengekstraksi data yang ditargetkan secara otomatis dari sebuah situs. Kemampuannya untuk mengumpulkan informasi dari sebuah situs dalam skala yang besar (Meschenmoser et al., 2016) menjadi teknik yang cocok untuk digunakan dalam penelitian ini, oleh karena perangkat lunak akan mengumpulkan data yang berjumlah besar dari berbagai situs riset.

Lalu, perangkat lunak akan memproses data yang telah dikumpulkan dengan model LDA, dengan luaran akhir visualisasi hasil pemodelan topik agar dapat digunakan oleh peneliti untuk memahami topik terkini dan tren perkembangan riset dalam topik apapun yang ditentukan oleh peneliti.

Tabel 2.1 merangkum dan menunjukkan ulasan kritis penulis terhadap setiap sub tema yang telah dibahas pada bab tinjauan pustaka ini. Pertama, penelitian secara umum dibahas, dimana tahap-tahap penelitian dapat diringkas dalam enam tahapan sekuensial yang direpresentasikan dalam metode ilmiah. Langkah awal utama yang krusial dalam melakukan penelitian adalah pengajuan pertanyaan dalam konteks pengetahuan yang ada, yang secara umum dikenal sebagai proses penemuan *state-of-the-art*.

State-of-the-art dapat diperoleh dari menemukan riset terkini pada sebuah bidang ilmu, agar dapat menunjukkan gambaran yang telah dilakukan dan yang dapat dilakukan berikutnya untuk mengembangkan bidang ilmu pengetahuan tertentu. Adapun beberapa cara untuk menemukan ini, yaitu menggunakan tinjauan pustaka, analisis bibliometris, dan pemodelan topik. Tinjauan pustaka merupakan salah satu cara yang melibatkan proses pengumpulan, analisis, dan sintesis temuan dari literatur. Namun, metode ini memiliki

kendala, seperti proses yang memakan waktu yang lama dan kesulitan untuk mengikuti penelitian yang mutakhir dalam topik penelitian tertentu. Analisis bibliometris telah digunakan sebagai metode alternatif untuk menemukan *research gap* dan topik terkini dari bidang ilmu yang ditentukan, namun masih memerlukan proses pengumpulan dan pemrosesan data manual dengan teliti dan hati-hati. Selain itu, pendekatan ini masih rentan terhadap kesalahan karena sifatnya yang terutama berfokus pada analisis berbasis kata kunci untuk konten artikel. Kemudian, pendekatan pemodelan topik seperti LDA merupakan metode pemodelan statistik yang digunakan untuk menemukan topik dalam dokumen dengan mengidentifikasi kumpulan kata yang terkait. Pendekatan ini telah digunakan sebelumnya dalam proses penemuan *state-of-the-art* juga, namun memiliki keterbatasan yang serupa dengan pendekatan yang sebelumnya, yang memerlukan keahlian serta proses yang lama dalam pengumpulan, pemrosesan, dan pemodelan data.

Tabel 2.1 Ringkasan dan Ulasan Kritis Pada Tinjauan Pustaka.

No.	Sub Tema	Keywords	Ringkasan dan Ulasan Kritis	Pustaka
1	Tahapan Penelitian	metode ilmiah, langkah-langkah penelitian	Proses penelitian dapat diringkaskan dalam enam tahapan sekuensial yang direpresentasikan dalam metode ilmiah. Langkah awal utama yang krusial dalam melakukan penelitian adalah pengajuan pertanyaan dalam konteks pengetahuan yang ada.	(Dodig Crnkovic, 2002; Santos, 2019)
2	Penemuan <i>state-of-the-art</i>	penemuan terkini, tren penelitian	Konteks pengetahuan yang ada dapat diperoleh dari menemukan riset terkini pada sebuah bidang ilmu, agar dapat menunjukkan gambaran yang telah dilakukan dan yang dapat dilakukan berikutnya untuk mengembangkan bidang ilmu pengetahuan tertentu.	(Jahan et al., 2016; Levy & Ellis, 2006; Santos, 2019; Snyder, 2019)

No.	Sub Tema	Keywords	Ringkasan dan Ulasan Kritis	Pustaka
3	Tinjauan pustaka	analisis dokumen, penggabungan ide, tinjauan pustaka sistematis	<p>Tinjauan pustaka merupakan salah satu cara untuk menemukan <i>state-of-the-art</i> pada sebuah topik, dengan mengumpulkan, menganalisis, dan mensintesis temuan dari literatur. Tinjauan pustaka sistematis, salah satu metode tinjauan pustaka yang dijadikan standar dalam melakukan tinjauan pustaka, merupakan sebuah cara untuk meninjau literatur dalam cara yang sistematis dan <i>reproducible</i>. Tinjauan pustaka mencakup tiga tahapan besar, yakni perencanaan, eksekusi, dan pelaporan. Tetapi metode penemuan <i>state-of-the-art</i> ini semakin sulit untuk dilakukan, oleh karena produksi ilmu yang semakin cepat dan proses yang membutuhkan waktu yang lama. Hal ini membuat sulit untuk mengikuti penelitian yang mutakhir dalam topik penelitian tertentu yang dapat memperlambat proses penelitian.</p>	(Davis et al., 2014; Felizardo et al., 2012; Jahan et al., 2016; Levy & Ellis, 2006; Snyder, 2019; Wahono, 2015)
4	Analisis Bibliometris untuk Menemukan <i>State-of-the-</i>	metode bibliometris, analisis bibliometris, <i>research gap</i> ,	<p>Analisis bibliometris menggunakan perangkat lunak <i>VOSViewer</i> dan atau <i>CitNet Explorer</i> merupakan salah satu teknik yang telah digunakan oleh berbagai peneliti untuk</p>	(Garfield, 1955; Lee et al., 2021; Oliveira et al., 2019)

No.	Sub Tema	Keywords	Ringkasan dan Ulasan Kritis	Pustaka
	<i>art</i> dalam Literatur	tren dalam literatur	menyimpulkan <i>research gap</i> dan tren dalam topik yang ditentukan berdasarkan hasil analisis yang dilakukan. Metode yang digunakan masih membutuhkan proses pengumpulan dan pemrosesan data manual dengan teliti dan hati-hati agar menghasilkan analisis yang baik. Selain itu, hasil analisis bibliometrik belum mampu mewakili topik-topik pada topik penelitian yang ditentukan, karena hanya menggunakan data bibliometrik.	
5	Pemodelan topik	metode alternatif analisis dokumen, penemuan topik	Pemodelan topik merupakan metode pemodelan statistik yang dapat digunakan untuk menemukan topik-topik laten dari kumpulan dokumen. Pemodelan topik menjadi alternatif dari analisis bibliometrik untuk memahami literatur ilmiah, karena kemampuannya untuk menemukan topik laten dari konten dalam sebuah kumpulan dokumen yang besar dan tidak terstruktur. Pemodelan topik sudah digunakan untuk mengkarakterisasi bidang ilmu dan menganalisis tren	(Boyack et al., 2011; Felizardo et al., 2012; Gläser et al., 2017; Griffiths & Steyvers, 2004; Lee et al., 2021; Nikolenko et al., 2015; Rumbut et al., 2020)

No.	Sub Tema	Keywords	Ringkasan dan Ulasan Kritis	Pustaka
			perkembangan topik pada topik penelitian tertentu.	
6	LDA	pemodelan topik, model probabilistik	LDA merupakan teknik populer untuk melakukan pemodelan topik pada sebuah korpus. Topik diasumsikan sebagai koleksi kata dari korpus, yang ketika digabungkan, merepresentasikan sebuah topik.	(Ballester & Penner, 2022; Bansal, 2019; Lee et al., 2021; Pascual, 2019)
7	LSA	pemodelan topik, faktorisasi matriks	LSA merupakan metode faktorisasi matriks yang merepresentasikan frekuensi <i>co-occurrence</i> kata dan dokumen dalam <i>high-dimensional space</i> , sehingga memungkinkan identifikasi topik laten dan telah digunakan untuk penemuan <i>state-of-the-art</i> .	(Landauer & Dutnais, 1997; Valdez et al., 2018)
8	Analisis Pemodelan Topik untuk Menemukan <i>State-of-the-art</i> dalam Literatur	analisis pemodelan topik, penemuan topik dominan pada penelitian	Dalam penemuan <i>state-of-the-art</i> dalam literatur, pemodelan topik telah digunakan oleh berbagai peneliti untuk membantu menemukan topik-topik dominan serta tren topik dalam topik penelitian yang ditentukan. Metode yang digunakan hanya terbatas pada satu basis data dan masih membutuhkan proses ekstraksi, skrining, dan pemrosesan data manual.	(Al-Ramahi et al., 2016; Lee et al., 2021; Li et al., 2019; Rumbut et al., 2020)

Secara keseluruhan, dalam tinjauan pustaka ini, penulis menunjukkan bahwa ada berbagai metode yang dapat digunakan untuk menemukan *state-of-the-art* dalam literatur,

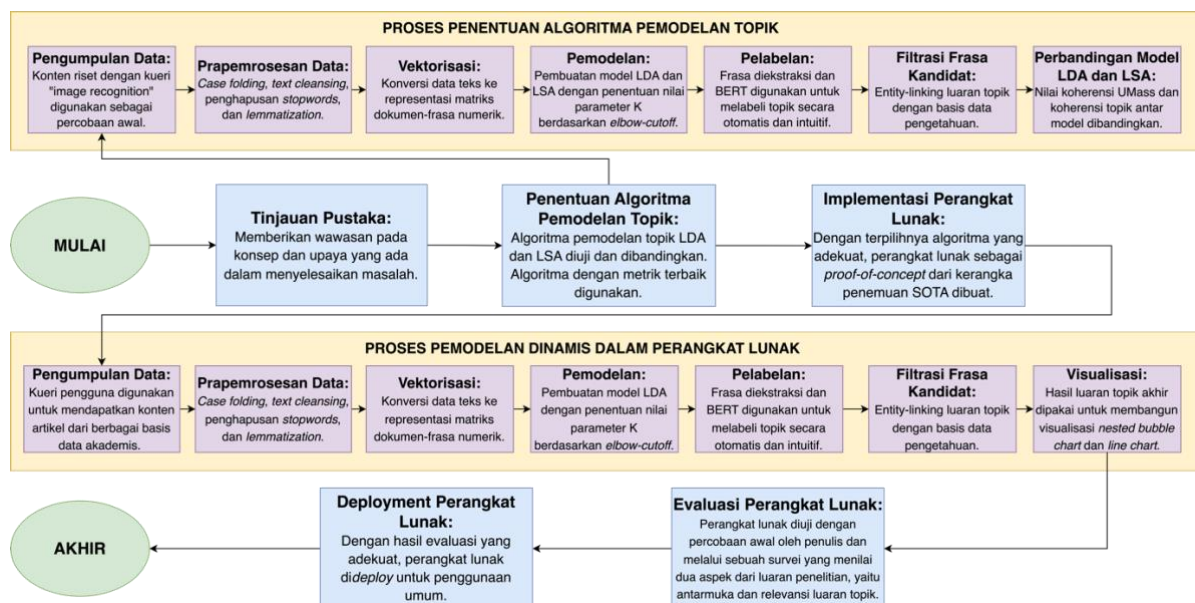
namun setiap metode memiliki kelebihan dan kekurangannya sendiri. Tabel 2.1 memberikan ringkasan, wawasan, dan ulasan kritis dari setiap pendekatan yang pernah ada, sehingga menjadi dasar dari pengembangan usulan pendekatan sebelumnya pada penelitian ini.

BAB 3

Metodologi

3.1 Langkah Penelitian

Gambar 3.1 menunjukkan sebuah diagram alir yang memvisualisasikan langkah-langkah yang diperlukan untuk melaksanakan penelitian. Adapun lima langkah utama yang dilakukan secara sekuensial dalam penelitian ini, yakni tinjauan pustaka, penentuan algoritma pemodelan topik, pembuatan perangkat lunak, evaluasi perangkat lunak, dan *deployment* perangkat lunak. Khususnya untuk langkah kedua dan ketiga, terdapat beberapa sub-proses yang dilakukan pada langkah-langkah tersebut. Hal ini dilambangkan dengan balok berwarna kuning yang mengenkapsulasi sub-proses (ditandai dengan warna ungu). Deskripsi lebih rinci mengenai masing-masing langkah dan sub-proses akan dijelaskan pada sub-bab berikutnya.



Gambar 3.1 Metode penelitian.

3.2 Tinjauan Pustaka

Penelitian dimulai terlebih dahulu dengan tinjauan pustaka, yang merupakan hal penting untuk dilakukan untuk memberikan pemahaman yang lebih dalam terhadap konsep dan upaya peneliti sebelumnya untuk menyelesaikan masalah penemuan *state-of-the-art*. Melalui tinjauan pustaka ini, peneliti dapat mengidentifikasi kelemahan dan kekurangan dari penelitian sebelumnya dan merumuskan solusi luaran penelitian yang sesuai dengan konteks

yang baru. Adapun penjelasan tentang konsep, wawasan, serta upaya peneliti sebelumnya yang menjadi dasar penelitian ini telah dipaparkan sebelumnya pada Bab 2.

3.3 Penentuan Algoritma Pemodelan Topik

Pemilihan algoritma merupakan langkah penting untuk penemuan *state-of-the-art* dalam penelitian ini. Dua algoritma pemodelan topik yang paling umum digunakan, yaitu LDA dan LSA, dipilih untuk dibandingkan. Alasan pemilihan kedua algoritma ini didasarkan pada popularitasnya dan penggunaan yang luas dalam penelitian sebelumnya.

LDA dan LSA semakin populer dalam beberapa tahun terakhir karena kemampuannya untuk mengekstraksi topik-topik yang bermakna dari kumpulan data teks yang besar dan tidak terstruktur, dan terbukti mengungguli algoritma pemodelan topik lainnya dalam beberapa penelitian. Penelitian sebelumnya telah membandingkan LDA dan LSA dalam berbagai bidang, termasuk pemrosesan bahasa alami, penemuan informasi, dan *state-of-the-art* (Bellaouar et al., 2021; Hassan et al., 2020; Suominen & Toivanen, 2016). Misalnya, Blei et al. (2003) membandingkan LDA dengan model lain, dan menunjukkan kemampuannya dalam mengungkap struktur yang mendasari dari korpus. Studi lain menunjukkan bahwa LSA terutama efektif untuk pemodelan topik pada dokumen dengan konten yang lebih pendek, sedangkan LDA lebih baik pada dokumen yang kontennya lebih banyak (Blei et al., 2003; Landauer & Dutnais, 1997). Selain itu, LSA dan LDA telah banyak diterapkan dalam penelitian yang menemukan dan menganalisis peta penelitian dan topik-topik dari sebuah bidang ilmu, yang tujuannya adalah untuk mengidentifikasi tren dan kesenjangan saat ini dalam suatu bidang penelitian, seperti yang dipaparkan sebelumnya pada tinjauan pustaka.

Mengingat popularitas dan penggunaan yang luas dari LSA dan LDA dalam penelitian sebelumnya, serta efektivitasnya dalam penemuan topik *state-of-the-art*, penelitian ini membandingkan kedua algoritma tersebut. Perbandingan bertujuan untuk mengidentifikasi algoritma yang paling sesuai untuk penemuan *state-of-the-art* secara otomatis, berdasarkan kemampuan untuk mengungkap topik-topik yang relevan dan memberikan label yang bermakna.

Maka dari itu, setelah tinjauan pustaka dilakukan, algoritma pemodelan topik LDA dan LSA diimplementasikan dan performanya dievaluasi pada basis data tolok ukur untuk menentukan algoritma pemodelan topik yang akan digunakan. Metrik yang akan digunakan untuk mengukur performa model adalah metrik koherensi *UMass* (Lee et al., 2021; Röder et al., 2015; Smatana & Butka, 2019) dengan data yang diperoleh dari basis data akademis

CORE untuk sebuah topik riset. Model dengan nilai metrik koherensi terbaik dan luaran topik yang koheren digunakan. Adapun proses yang akan diikuti dijabarkan pada sub-bagian selanjutnya.

3.3.1 Pengumpulan Data

Pengumpulan data untuk penentuan algoritma dilakukan dengan mengumpulkan data riset, atau, lebih spesifiknya data abstrak dari riset dalam bidang “*image recognition*” pada basis data akademis CORE. Hal ini dilakukan menggunakan bahasa pemrograman Python dengan pustaka *requests*, dimana pustaka tersebut digunakan untuk membuat permintaan *hypertext transfer protocol* (HTTP) ke *application programming interface* (API) yang disediakan oleh CORE, dengan kueri pencarian riset sebagai berikut: “*image recognition*” AND *fieldsOfStudy*:“*computer science*” AND *documentType*:“*research*”. Kueri ini akan mengembalikan semua artikel akademis dalam bidang ilmu komputer yang mengandung topik *image recognition*.

Dari hasil data yang diperoleh, hanya data abstrak yang akan digunakan sesuai dengan yang dilakukan pada penelitian sebelumnya yang menganjurkan penggunaan abstrak atas dasar ketersediaan data dan gaya penulisan yang konsisten (Daenekindt & Huisman, 2020; Griffiths & Steyvers, 2004). Selain itu, pengumpulan data untuk tahap penentuan algoritma ditetapkan pada 500 artikel saja. Adapun kode yang digunakan untuk melakukan hal ini dapat dilihat di <https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py>. Data yang diperoleh dari basis data CORE menggunakan kode diatas disimpan dalam bentuk CSV menggunakan pustaka *pandas*.

3.3.2 Prapemrosesan Data

Sebelum menggunakan data, prapemrosesan dilakukan terlebih dahulu pada data agar memastikan data layak digunakan untuk pemodelan (Holzinger, 2019). Tahapan-tahapan prapemrosesan dilakukan dengan menggunakan pustaka *gensim*, dan keseluruhan proses dapat dilihat pada <https://github.com/timothydillan/ldaviewer/blob/main/api/lda/ldapipeline.py#L66>.

Penjelasan detail mengenai proses prapemrosesan yang dilakukan adalah sebagai berikut:

a) *Case Folding*

Case folding, dengan membuat semua teks huruf kecil, dilakukan untuk memastikan bahwa teks distandarisasi dalam bentuk *casenya*. Proses ini dilakukan untuk memastikan bahwa model pembelajaran mesin menafsirkan kata-kata dengan *case* karakter yang berbeda (misalnya, "Anda" dan "ANDA") dengan cara yang sama, meminimalkan kemungkinan model akan menimbang kata-kata yang tepat secara berbeda (Arminditya Fajri Akbar et al., 2021; Resti & Selva Jumeilah, 2017).

b) *Text Cleansing*

Text cleansing lalu dilakukan yang mencakup beberapa subproses, seperti normalisasi aksan dan diakritik, di mana kata-kata seperti "*cliché*" diubah menjadi "*cliche*". Selain itu, penghapusan tautan, tanda baca, angka, dan karakter non-Latin juga dilakukan (Arminditya Fajri Akbar et al., 2021; Resti & Selva Jumeilah, 2017).

c) *Penghapusan Stopwords*

Penghapusan *stopwords* juga dilakukan untuk membuang kata-kata yang sering muncul dalam teks, tapi tidak berkontribusi pada arti teks (Arminditya Fajri Akbar et al., 2021; Resti & Selva Jumeilah, 2017).

d) *Lemmatization*

Lemmatization, sebagai tahap terakhir prapemrosesan dilakukan untuk mereduksi kata menjadi kata dasarnya. Hal ini bertujuan untuk memastikan bahwa kata-kata turunan dapat diartikan dengan cara yang sama oleh model (Arminditya Fajri Akbar et al., 2021; Resti & Selva Jumeilah, 2017).

3.3.3 Vektorisasi

Dengan data yang telah dipraproses, data diubah menjadi representasi angkanya terlebih dahulu sebelum dapat digunakan dalam pemodelan. Sebagian besar implementasi pemodelan topik melakukan hal ini dengan membangun sebuah matriks frekuensi untuk setiap token, dimana token berupa sebuah *sequence* *n*-gram, dimana *n* ditentukan oleh peneliti. Implementasi tersebut tidak dilakukan pada penelitian ini, dikarenakan beberapa alasan, diantaranya penentuan *n*-gram membutuhkan peneliti untuk mengetahui domain agar dapat menentukan *n*-gram yang sesuai.

Selain itu, satu konfigurasi *n*-gram yang telah digunakan dapat tidak berlaku untuk kasus-kasus lainnya, sehingga penentuan *n*-gram membutuhkan percobaan berulang kali. Hal ini dapat menyebabkan hasil luaran model mengeluarkan frasa-frasa yang “canggung” dan tidak benar secara tata bahasa (Ali et al., 2015; Rabby et al., 2018; Sidorov et al., 2012). Maka dari itu, penelitian ini menggunakan pendekatan yang dapat mengekstraksi *sequence*

atau frasa menggunakan pola *part-of-speech* (POS). Adapun implementasi *vectorizer* yang digunakan dapat dilihat pada <https://github.com/timothydillan/ldaviewer/blob/main/api/lda/ldapipeline.py#L158>. Pola POS yang digunakan adalah sebagai berikut:

```
{<J.*><N.*><N.*>|<J.*><N.*>|<J.*><N.*>+}
{<N.*><N.*>}
{<N.*><IN.*><N.*>}
```

Pola POS yang digunakan adalah pola yang banyak digunakan dalam penelitian untuk mengekstraksi frasa, dimana semua frasa yang dimulasi dengan sebuah *adjective* (kata sifat) yang diikuti oleh *noun* (kata benda), atau *noun* dengan *noun*, atau *noun* dengan *preposition* (kata preposisi) dan *noun* diekstraksi (Z. Liu et al., 2009; Mihalcea & Tarau, 2004; Wan & Xiao, 2008; Zervanou, 2010). Selain itu, terdapat filtrasi tambahan pada proses vektorisasi data, dimana frasa atau kata-kata yang muncul di lebih dari 93% dokumen dalam korpus akan dibuang. Dengan pendekatan ini, tidak lagi diperlukan definisi n-gram secara manual dan hasil pemodelan dapat menghasilkan luaran yang diharapkan lebih koheren dan secara tata bahasa, benar.

3.3.4 Pemodelan

Representasi data teks yang sudah ditransformasi ke dalam bentuk numerik lalu digunakan untuk membangun model LDA dan LSA dengan bantuan pustaka *sklearn* dan *numpy* (Harris et al., 2020; Pedregosa et al., 2011). Pada pembuatan model LDA, beberapa parameter perlu ditetapkan secara manual, yakni parameter: α , β , and K . Penentuan nilai parameter α dan β dilakukan secara otomatis dengan metode *variational Bayes inference* yang disediakan oleh pustaka *sklearn*. Nilai parameter K ditentukan menggunakan metode *elbow cut-off*, dimana nilai parameter K dengan skor koherensi *intrinsic UMass* tertinggi digunakan. Mekanisme pencarian nilai parameter K dapat dilihat pada potongan kode di <https://github.com/timothydillan/ldaviewer/blob/main/api/lda/ldapipeline.py#L324>. Mirip dengan LDA, nilai parameter K pada LSA ditentukan dengan cara yang serupa dengan LDA, yakni menggunakan metode *elbow cut-off*.

3.3.5 Pelabelan Topik

Hasil pemodelan topik yang dilakukan mengembalikan topik, kata-kata yang merepresentasikan topik tersebut, dan bobot setiap kata, lalu dianalisis lebih lanjut dengan

mengeksktraksi frasa-frasa yang secara semantik dapat mewakili topik dengan baik. Ini dilakukan dengan menggabungkan 25 kata dengan bobot terbesar dari setiap topik dan lalu menghasilkan frasa-frasa kandidat dengan cara yang serupa dengan vektorisasi pada pemomdelan. Model BERT digunakan untuk membuat representasi *embedding* dari penggabungan dan setiap kandidat. Lalu, fungsi *cosine similarity* diaplikasikan pada representasi *embedding* setiap kandidat dan penggabungan kata (Koloski et al., 2022). Lima kandidat frasa yang memiliki nilai kesamaan yang paling tinggi akan digunakan. Proses ini dilakukan dengan menggunakan pustaka *keybert* dan *keyphrase_vectorizer* yang implementasinya dapat dilihat di <https://github.com/timothydillan/ldaviewer/blob/main/api/lda/ldapipeline.py#L501>.

3.3.6 Filtrasi Frasa Kandidat

Frasa-frasa yang dihasilkan dapat saja menghasilkan frasa yang “acak”, atau tidak merepresentasikan suatu konsep atau topik asli. Maka dari itu, frasa-frasa difiltrasi kembali dengan melakukan *entity linking*, yang merupakan proses dimana setiap frasa di *resolve* ke *knowledge base* (KB) seperti ConceptNet dan Wikipedia (Hachey et al., 2013; Han et al., 2021; Shen et al., 2015). Frasa-frasa yang tidak ditemukan pada KB akan dibuang.

ConceptNet merupakan sebuah basis data ilmu yang dibangun oleh MIT. Basis data, atau grafik pengetahuan ini mengandung informasi tentang sebuah kata atau frasa dan hubungan data tersebut dengan data lainnya (Speer et al., 2017). Mirip dengan ConceptNet, Wikipedia juga merupakan sebuah basis data atau ensiklopedia yang mengandung informasi tentang semua cabang ilmu (Medelyan et al., 2009). Asumsi dasarnya disini adalah karena ConceptNet dan Wikipedia mencakup konsep-konsep dan hubungan yang sudah diverifikasi oleh manusia, maka jika frasa yang digunakan tidak ada dalam kedua platform tersebut, frasa tersebut diasumsikan tidak ada dan dibuang. Implementasi filtrasi frasa menggunakan ConceptNet dan Wikipedia dapat dilihat pada <https://github.com/timothydillan/ldaviewer/blob/main/api/lda/ldapipeline.py#L654>. Pada dasarnya, setiap frasa akan digunakan sebagai kueri pencarian pada KB. Jika frasa tidak ditemukan pada KB, dan bila frasa memiliki lebih dari satu kata, maka program akan mengkonstruksi kombinasi n-gram dari setiap kata dalam frasa untuk $N = 2$ hingga jumlah kata pada frasa dan mencarinya pada KB. Jika salah satu kombinasi kata ditemukan, maka frasa tetap digunakan. Adapun pustaka tambahan yang digunakan dalam proses ini yakni pustaka *conceptnet_lite* sebagai antarmuka ke basis data ConceptNet dan *wikipedia* untuk mengakses data Wikipedia melalui permintaan HTTP.

3.3.7 Perbandingan Model LDA dan LSA

Berdasarkan luaran nilai metrik koherensi *UMass* dan topik-topik yang dihasilkan oleh kedua model LDA dan LSA pada data yang diberikan, kedua model akan dibandingkan dan penentuan algoritma pemodelan topik yang akan digunakan didasari pada nilai koherensi yang lebih baik dan luaran topik yang koheren dan tersegregasi.

3.4 Implementasi Perangkat Lunak

Setelah model telah ditentukan, sebuah perangkat lunak yang dapat melakukan pengumpulan data, prapemrosesan data, dan pemodelan topik berdasarkan model dibangun. Sub-bab berikut menjelaskan secara detil masing-masing tahapan yang dilakukan dalam implementasi perangkat lunak.

3.5 Pengumpulan Data

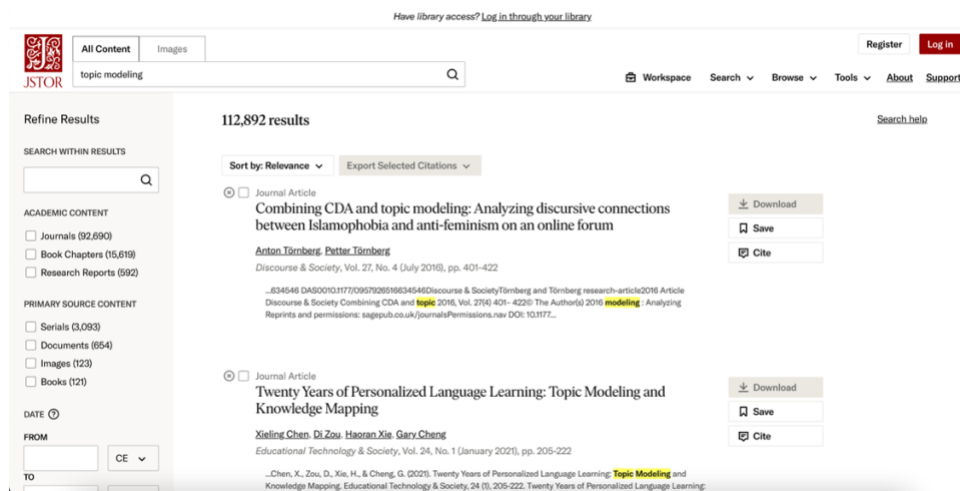
Perangkat lunak dapat secara otomatis mengumpulkan data artikel dari kata kunci pencarian yang diberikan, dan data yang diperoleh akan mengandung data bibliometrik, seperti judul artikel, nama penulis, dan konten artikel. Pengumpulan data pada luaran riset ini menggunakan lima basis data akademis terbuka, yaitu basis data CORE, arXiv, Emerald, ScienceOpen, dan Garuda.

Semua implementasi pemerolehan data dari basis data yang digunakan dapat dilihat pada

<https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py>.

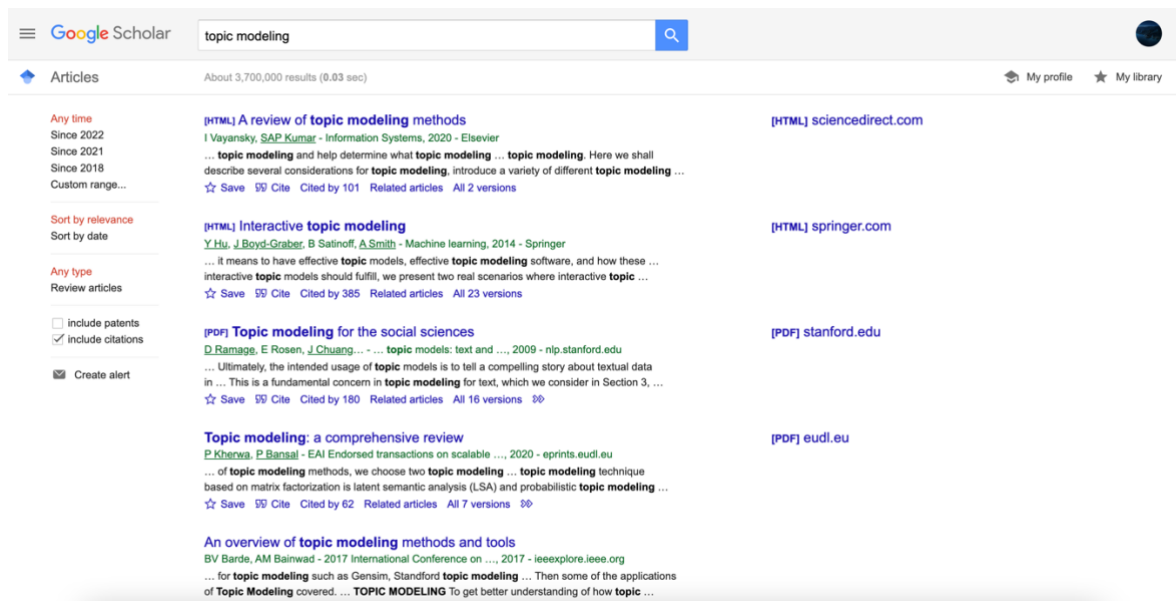
Beberapa basis data yang telah dipaparkan pada proposal penelitian ini, seperti Scopus, JSTOR, dan Google Scholar, tidak digunakan dikarenakan beberapa pertimbangan teknis dan konten. Basis data Scopus, atau ScienceDirect, memberlakukan pembatasan akses pada konten akademis jika pengguna tidak memiliki akun atau akses *proxy* institusional (lihat Gambar 3.2).

membuat permintaan HTTP yang dapat berjumlah sangat banyak, juga ditambah dengan kompleksitas untuk menekan tombol penunjukkan abstrak secara otomatis agar dapat memicu permintaan HTTP. Maka dari itu, basis data ScienceDirect tidak digunakan. Selain ScienceDirect, JSTOR juga tidak digunakan atas dasar beberapa alasan teknis. Ketika mengakses JSTOR, data yang bisa didapatkan hanyalah judul, penulis, dan *excerpt* atau kutipan artikel (lihat Gambar 3.4).



Gambar 3.4 Pencarian artikel pada basis data JSTOR.

Dikarenakan konten yang dapat diperoleh melalui *scraping* hanyalah *excerpt* atau kutipan dari artikel yang mengandung kata kunci pencarian, dan karena data ini hanya menyorot satu kalimat yang mengandung kata kunci yang dicari, konten yang diperoleh dapat menjadi tidak relevan. Jika diaplikasikan pemodelan topik pada data *excerpt* dapat berakibat pada luaran topik yang kurang sesuai dikarenakan teks yang tidak menggambarkan keseluruhan konten dari topik yang ingin dicari, sehingga basis data ini tidak digunakan. Untuk alasan yang sama, Google Scholar juga tidak digunakan dikarenakan hanya dapat menunjukkan *excerpt* artikel (lihat Gambar 3.5).



Gambar 3.5 Pencarian artikel pada basis data Google Scholar.

Selain itu, Google Scholar juga mengimplementasikan berbagai fitur anti-bot seperti pengecekan reCAPTCHA yang dapat menghambat peroleh informasi secara otomatis. Maka dari itu basis data Google Scholar juga tidak digunakan.

3.5.1 CORE

Berbeda dengan basis data diatas, basis data CORE menyediakan sebuah API yang dapat dengan mudah digunakan untuk mendapatkan data yang disediakan oleh CORE. CORE adalah agregator makalah penelitian *open-access* terbesar di dunia dari repositori dan jurnal yang dibangun oleh Open University dan Jisc (CORE, n.d.-a). Adapun data-data yang disediakan pada API CORE, seperti daftar jurnal yang ada pada basis data dan konten riset (judul, abstrak, *full-text*, penulis artikel) (CORE, n.d.-b). APInya pun disediakan secara gratis untuk penggunaan non-komersil. Maka, basis data CORE merupakan salah satu pilihan basis data yang digunakan dalam penelitian.

Proses pengambilan data pada API diatas didasarkan pada permintaan HTTP ke situs API yang dikhususkan untuk pengambilan data riset. Permintaan diisi dengan kueri pencarian pengguna, dengan limit pengambilan data sebesar 10,000 data dalam satu permintaan (batasan yang dikenakan dari basis data). Adapun data yang diambil berupa data abstrak dan *full-text*. Implementasi pengambilan data pada basis data CORE dapat dilihat di: <https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py#L119>.

3.5.2 arXiv

Selain CORE, basis data lain yang juga menjadi opsi untuk digunakan adalah basis data arXiv. Dibangun oleh Cornell University, arXiv merupakan basis data akademis terbuka yang mengandung lebih dari dua juta artikel ilmiah di bidang fisika, matematika, ilmu komputer, biologi kuantitatif, keuangan kuantitatif, statistik, teknik elektro, ilmu sistem, dan ekonomi. arXiv, seperti CORE, menyediakan API yang dapat memberikan akses untuk mengumpulkan data secara otomatis melalui kueri pencarian (arXiv, 2021).

Pustaka arXiv.py digunakan untuk pengambilan data dari arXiv (Schwab, 2021). Juga mirip dengan cara kerja pengambilan pada basis data CORE, pustaka arXiv.py melakukan pengambilan data menggunakan permintaan HTTP dari kueri pencarian pengguna, dimana proses pengambilan data ini dibungkus dalam fungsi-fungsi yang dapat digunakan dengan mudah oleh pengembang. Kode untuk melakukan pengambilan data pada basis data arXiv dapat dilihat di: <https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py#L529>.

3.5.3 Emerald

Emerald, atau Emerald Insights, merupakan sebuah basis data yang mengandung artikel yang diperoleh dari lebih dari 300 *peer-reviewed* jurnal, yang mencakup bidang pemasaran, manajemen, sumber daya manusia, kualitas, ekonomi, manajemen informasi, operasi, *engineering*, produksi dan properti, pendidikan, kesehatan dan perawatan sosial (University of Cambridge, n.d.; University of the West of Scotland, 2022).

Pengambilan data pada Emerald dilakukan melalui *browser automation* dengan pustaka selenium dikarenakan Emerald tidak menyediakan sebuah antarmuka teknis seperti API untuk pengambilan data. Dengan selenium, setelah situs Emerald Insights dibuka dengan kueri pencarian pengguna, semua abstrak yang ada akan diambil secara otomatis dengan mendapatkan semua teks yang ada di dalam elemen yang mewadahi abstrak. Adapun batasan dalam kemampuan *scraping* pada basis data ini, yakni hanya mendapatkan konten abstrak untuk riset dalam bentuk artikel ilmiah saja dari kueri pencarian pengguna. Elemen-elemen dan atribut *hypertext markup language* (HTML) yang digunakan dalam proses *scraping* dengan selenium dapat dilihat di: <https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py#L355>.

3.5.4 ScienceOpen

ScienceOpen adalah basis data akademis yang mengandung lebih dari 80 juta artikel terbuka yang diperoleh dari 25 ribu jurnal dan 30 juta penulis dan mencakup semua bidang ilmu. Selain menjadi basis data, ScienceOpen juga menjadi platform bagi peneliti untuk mempublikasikan riset dengan pengguna ScienceOpen lainnya sebagai *peer-reviewer*. Untuk jurnal atau penerbit, ScienceOpen dapat menjadi platform untuk menampilkan dan menyimpan riset (ScienceOpen, 2017).

Mirip dengan Emerald, ScienceOpen tidak menyediakan sebuah API untuk pengambilan data. Maka dari itu, *browser automation* dilakukan untuk melakukan *scraping* pada basis data ini. Saat ini, implementasi *scraping* pada ScienceOpen mendukung pencarian dari frasa, abstrak, dan filtrasi artikel melalui rentang tanggal yang ditentukan oleh pengguna. Adapun batasan serupa pada basis data Emerald yang diimplementasikan pada proses *scraping* ScienceOpen, dimana data yang diperoleh hanya berupa abstrak dari konten riset dengan jenis artikel ilmiah. Implementasi *scraping* pada basis data ini dapat dilihat di:

<https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py#L616>.

3.5.5 Garuda

Garba Rujukan Digital (Garuda) merupakan basis data karya ilmiah peneliti Indonesia yang dikembangkan oleh kementerian pendidikan, kebudayaan, riset, dan teknologi Indonesia. Adapun cakupan karya ilmiah yang terindeks oleh Garuda antara lain *e-journal* domestic, konferens, dan laporan penelitian, dengan lebih dari 2 juta artikel yang diperoleh dari 16 ribu jurnal, 215 konferens, dan tiga ribu penerbit dan mencakup semua bidang ilmu (Garuda - Garba Rujukan Digital, 2018).

Basis data Garuda juga tidak menyediakan sebuah API untuk melakukan pengambilan data. Maka, *browser automation* dilakukan untuk melakukan *scraping* pada basis data ini juga. Adapun batasan serupa pada implementasi proses *scraping* pada Garuda dimana data yang diperoleh hanya berupa abstrak, dan filtrasi kata kunci pencarian didasarkan pada keberadaanya di dalam abstrak karya ilmiah yang disediakan di Garuda. Implementasi *scraping* pada basis data Garuda dapat dilihat pada tautan berikut: <https://github.com/timothydillan/ldaviewer/blob/main/api/papersearch/abstractscraper.py#L786>.

3.6 Pemodelan pada Runtime

Dari semua basis data yang akan digunakan, luaran penelitian ini dapat mengagregasikan semua data yang telah dikumpulkan, melakukan pemodelan topik abstrak setiap artikel, dan menunjukkan tren dan *state-of-the-art* dari kueri pencarian penggunaan. Fokus pemodelan topik pada abstrak artikel ini sesuai dengan studi-studi sebelumnya yang didukung oleh beberapa alasan lainnya seperti: (i) ketersediaan data abstrak dari basis data yang digunakan, (ii) abstrak mewakili ringkasan singkat artikel, yang meminimalkan kemungkinan mengidentifikasi topik minor, dan (iii) abstrak antar artikel cukup sebanding dalam hal format dan gaya penulisan (Daenekindt & Huisman, 2020; Griffiths & Steyvers, 2004). Proses pemodelan ini akan dilakukan pada *runtime*, dengan maksud model dibuat secara dinamis dengan data abstrak dari hasil pencarian pengguna. Tahapan-tahapan yang dilakukan untuk pemodelan dibuat dalam suatu *pipeline*, dengan langkah sebagai berikut.

3.6.1 Prapemrosesan Data

Pada tahap prapemrosesan, serupa dengan yang telah dipaparkan pada bab eksperimen awal dan metodologi, mencakup proses *case folding*, *text cleansing*, penghapusan *stopwords*, dan *lemmatization*. Karena pengumpulan data melibatkan pengumpulan abstrak artikel yang berbahasa Indonesia dan Inggris, maka daftar *stopwords* yang digunakan merupakan gabungan dari *stopwords* Indonesia dan Inggris yang disediakan oleh pustaka NLTK (Bird et al., 2009). Selain itu, untuk *lemmatization*, pustaka *stanfordnlp* dengan *pipeline* NLP bahasa Indonesianya digunakan untuk menemukan *lemma* dari teks berbahasa Indonesia (Manning et al., 2014), dan *spacy* digunakan untuk teks berbahasa Inggris (Montani et al., 2022).

3.6.2 Vektorisasi

Setelah melakukan prapemrosesan, proses mengubah teks ke representasi numeriknya, atau vektorisasi, dilakukan dengan menggunakan *KeyphraseCountVectorizer* yang menghasilkan matriks dokumen-frasa (matriks yang menjelaskan frekuensi frasa dalam dokumen) menggunakan pendekatan POS. Seperti tahap prapemrosesan, deteksi POS untuk data berbahasa Indonesia menggunakan *pipeline* berbahasa Indonesia dari pustaka *stanfordnlp*, dan *spacy* digunakan untuk yang berbahasa Inggris.

3.6.3 Pemodelan

Matriks numerik yang telah diekstraksi pada tahap sebelumnya lalu digunakan untuk menentukan nilai parameter pada model yang akan dibangun. Adapun beberapa parameter yang akan dihardcode pada model, sebagai berikut:

- $max_iter = 50$
- $learning_method = batch$
- Parameter α dan β ditentukan melalui secara otomatis menggunakan *variational Bayes inference*.

Untuk parameter K atau jumlah topik akan ditentukan melalui penggunaan *elbow cut-off*, dimana nilai parameter K dengan skor koherensi *intrinsic UMass* tertinggi digunakan. Mekanisme penghentian awal atau *early stopping* juga diterapkan untuk penentuan nilai K , dengan $e = 8$, di mana e adalah parameter kesabaran (*patience parameter*). Penghentian awal ini mengikuti logika sebagai berikut: untuk $i \dots j$, dimana i dan j adalah nilai yang telah ditentukan (dalam penelitian ini, 2 hingga 30), proses pencarian nilai K dihentikan jika skor koherensi tidak meningkat pada $i \dots i+e$. Setelah mendapatkan nilai-nilai parameter yang dapat diasumsikan merupakan parameter yang optimal, pembuatan model LDA akan dilakukan dengan nilai parameter yang ditemukan.

3.6.4 Pelabelan Topik

Pelabelan topik secara otomatis lalu dilakukan dengan cara yang serupa yang telah dipaparkan pada ekseprimen awal, dimana 25 kata dengan bobot terbaik dari setiap topik akan digabungkan menjadi satu teks. Lalu, menggunakan proses ekstraksi frasa yang diakomodasi oleh model *deep learning pretrained* multilingual BERT *paraphrase-multilingual-MiniLM-L12-v2* (Reimers & Gurevych, 2019), 10 frasa yang paling merepresentasikan setiap topik akan ditemukan. Ini dilakukan dengan mencari frasa dengan kemiripan kosinus *embedding* terdekat pada gabungan 25 kata dari setiap topik. Diasumsikan bahwa kemiripan yang lebih tinggi mengindikasikan bahwa frasa-frasa tersebut dapat mewakili topik secara keseluruhan dengan baik. Adapun alasan mengapa model multilingual digunakan dikarenakan korpus yang diproses dalam pemodelan dapat mengandung data berbahasa Indonesia dan Inggris, dan juga masing-masing teks dapat mengandung kata, frasa, atau istilah dalam bahasa lainnya. Selain itu, model *pretrained* BERT yang digunakan telah *fine-tune* untuk mengidentifikasi dan menemukan kemiripan semantik antar frasa dalam bahasa yang berbeda, sehingga tepat untuk digunakan pada konten artikel yang mengandung berbagai bahasa.

Pemilihan metode pelabelan menggunakan model BERT juga dilakukan aspek kualitas *embedding* yang dihasilkan. Berbagai metode lainnya untuk menghasilkan *embedding*, seperti word2vec, GloVe, FastText, dan ELMO, umumnya tidak dilatih pada korpus teks yang besar, tidak menggunakan arsitektur transformer, dan juga tidak memperhitungkan konteks kata di dalam teks. BERT sebagai model dasar telah dilatih pada korpus teks yang sangat besar dan mencakup 104 bahasa, sehingga telah melihat banyak konteks yang berbeda, dan mampu menghasilkan *embedding* yang memperhitungkan konteks di mana kata muncul (Devlin et al., 2018; Geetha & Karthika Renuka, 2021). Selain itu, karena model yang digunakan adalah *pretrained* dan telah di *fine-tune* pada tugas penyocokan frasa, membuatnya mampu menghasilkan *embedding* yang lebih akurat dan berguna untuk tugas pengolahan bahasa alami dibandingkan dengan metode *embedding* lainnya.

3.6.5 Filtrasi Frasa Kandidat

Frasa-frasa kandidat yang diasumsikan dapat mewakili topik dengan baik lalu diproses lebih lanjut untuk memastikan bahwa masing-masing frasa merujuk ke sebuah topik, fenomena, atau konsep yang nyata. Metode serupa yang telah dilakukan sebelumnya pada tahap penentuan algoritma pemodelan topik digunakan, dimana masing-masing frasa akan diverifikasi dan difiltrasi kembali dengan basis data pengetahuan seperti ConceptNet dan Wikipedia. Frasa yang tidak ditemukan pada kedua basis data tersebut, akan difiltrasi atau dibuang, sehingga hanya frasa-frasa yang ditemukan saja yang akan dipakai untuk merepresentasikan topik.

3.7 Visualisasi

Luaran dari pemodelan, yakni frasa yang merepresentasikan setiap topik, beserta dengan bobotnya, akan digunakan dalam pemvisualisasian hasil akhir pemodelan topik. Setiap dokumen abstrak yang dikumpulkan akan didapatkan probabilitas dikategorikannya ke topik-topik tertentu dengan model yang telah dibangun. Lalu, setiap dokumen akan dikategorikan ke topik tertentu dengan nilai probabilitas topik tertinggi.

Visualisasi pertama yang dihasilkan adalah distribusi pengelompokan topik pada kueri pencarian pengguna menggunakan sebuah *bubble chart*. Berdasarkan jumlah dokumen, frasa representatif, dan bobot setiap frasa dari setiap topik, sebuah *bubble* akan dibuat dengan ukuran yang didasarkan pada jumlah dokumen yang dikategorikan topik tertentu, label dari *bubble* adalah frasa dengan bobot terbesar, dan terdapat *nested bubbles*

yang merepresentasikan frasa-frasa representatif lainnya (jika ada). *Nested bubble chart* ini diharapkan dapat menunjukkan kelompok topik yang sedang tren dari kata kunci yang ditentukan oleh pengguna.

Selain *bubble chart*, *line chart* juga dihasilkan untuk menunjukkan evolusi frekuensi topik dari waktu-ke-waktu berdasarkan kueri pencarian pengguna. Axis Y pada grafik akan merepresentasikan jumlah dokumen, dan axis X merepresentasikan rentang waktu. Setiap poin data pada grafik akan merepresentasikan jumlah dokumen untuk topik dan waktu tertentu.

3.8 Evaluasi Perangkat Lunak

Performa perangkat lunak yang telah dibangun dalam penemuan *state-of-the-art* lalu diuji terlebih dahulu sebelum perangkat lunak *dideploy* untuk penggunaan umum. Kata kunci pencarian yang digunakan pada tahap seleksi algoritma akan digunakan pada perangkat lunak yang telah dibangun, dan dievaluasi luaran topiknya secara subjektif oleh peneliti. Selain itu, perangkat lunak yang telah dibangun lalu selanjutnya diuji lebih lanjut untuk memastikan bahwa secara kualitatif, perangkat lunak layak untuk digunakan dalam penelitian dan secara umum oleh publik. Proses evaluasi kualitatif ini meliputi sebuah kuesioner yang telah dibuat untuk menilai dua dimensi utama dari luaran penelitian, yakni aplikasi dan hasil pemodelan topik.

Kuesioner ini disebarkan ke individu yang merupakan dosen dan atau peneliti mahasiswa. Adapun alasan mengapa individu dengan kualifikasi tersebut dipilih, untuk memastikan bahwa responden memiliki gagasan yang jelas tentang *state-of-the-art* pada fokus riset responden masing-masing. Kriteria pemilihan responden ini juga digunakan untuk memastikan bahwa masing-masing responden kompeten dan dapat menilai relevansi luaran topik serta kesesuaiannya dengan bidang ilmu yang diminati. Selain itu, pada kuesioner telah diberikan informasi bahwa *state-of-the-art* yang dimaksud adalah *state-of-the-art* teoretis yang merujuk ke topik terkini dalam bidang ilmu.

Setiap dimensi memiliki tiga indikator yang akan digunakan untuk memberikan evaluasi secara keseluruhan terhadap masing-masing dimensi, dan dinilai dalam skala likert, dengan abreviasi Sangat Tidak Setuju (STS – Skor 1), Tidak Setuju (TS – Skor 2), Setuju dengan Keraguan (SK – Skor 3), Setuju (S – Skor 4), dan SS (Sangat Setuju – Skor 5). Adapun butir-butir yang digunakan dalam kuesioner dapat dilihat pada Tabel 3.1.

Tabel 3.1 Butir Kuesioner Evaluasi Perangkat Lunak.

No.	Dimensi	Indikator	Skala				
			STS	TS	SK	S	SS
A1.	Aplikasi	Antarmuka aplikasi ini mudah untuk dinavigasi					
A2.		Saya bisa mendapatkan informasi dengan mudah dari antarmuka aplikasi ini					
A3.		Informasi pada aplikasi ini sangat mudah dicerna					
PT1.	Pemodelan Topik	Dengan aplikasi ini, saya mendapatkan topik-topik yang relevan dengan topik yang saya pilih.					
PT2.		Artikel yang digunakan sesuai dengan label topik yang diberikan dari proses pemodelan topik.					
PT3.		Hasil luaran dan visualisasi model ini berguna untuk riset saya.					

Butir kuesioner yang dipaparkan pada Tabel 3.1 dibuat dalam sebuah formulir menggunakan *Google Form* yang digunakan untuk memudahkan proses pengumpulan evaluasi dari luaran perangkat lunak. Setiap partisipan lalu diminta untuk mencoba perangkat lunak yang telah dibangun, dan mengisi formulir yang telah diberikan sesuai dengan penilaian masing-masing partisipan. Jika hasil evaluasi tersebut menunjukkan bahwa secara rata-rata, nilai masing-masing butir untuk kedua dimensi diatas tiga, perangkat lunak diasumsikan layak untuk penggunaan publik.

Sampel luaran dari salah satu entri yang dimasukkan oleh responden juga dievaluasi dan dibandingkan dengan hasil dari perangkat lunak yang memiliki tujuan dan visualisasi yang serupa, yakni *Open Knowledge Maps*. *Open Knowledge Maps* merupakan perangkat lunak yang dikembangkan untuk membantu dalam pencarian, penemuan, dan pemahaman literatur ilmiah pada bidang ilmu. Sejalan dengan tujuan riset ini, *open knowledge maps* dibangun atas dasar sulitnya menemukan peta penelitian dari banyaknya konten akademis yang tersedia. Sehingga, perangkat lunak tersebut menyediakan sebuah antarmuka yang memvisualisasikan sebuah topik penelitian dalam bentuk peta pengetahuan, yang berisi

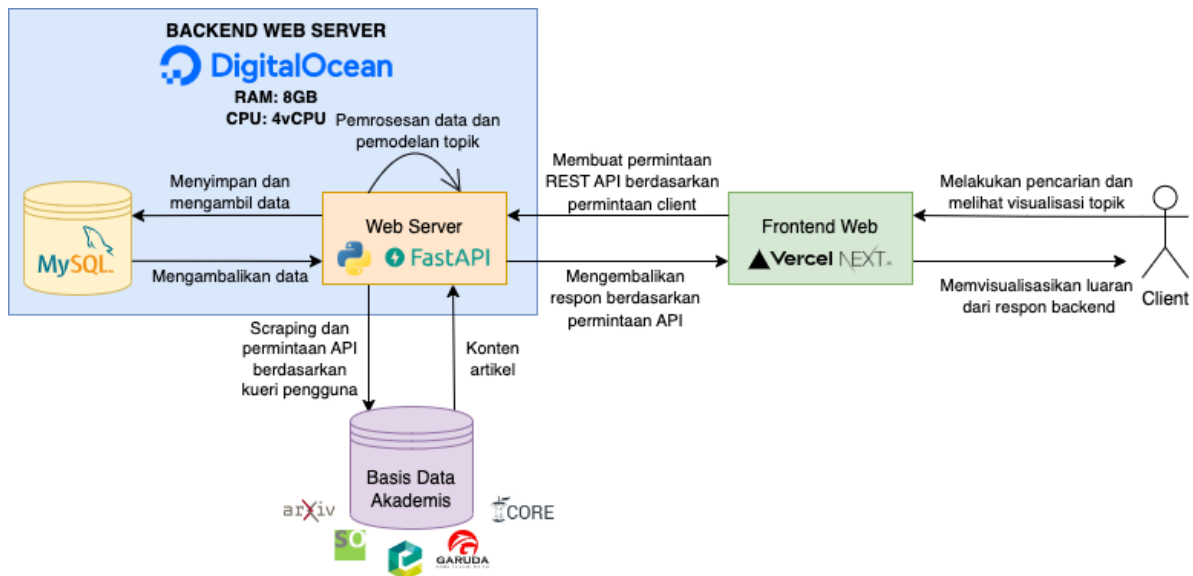
tema-tema utama dalam bidang ilmu yang direpresentasikan dalam bentuk lingkaran (Kraker et al., 2016, 2019; H. Liu et al., 2021).

Open Knowledge Maps telah banyak digunakan pada penelitian dari berbagai bidang untuk membantu peneliti dalam memahami peta penelitian, mengidentifikasi tren riset, serta menemukan kesenjangan dalam riset. Semisal, (Karampela et al., 2021; Mandaka et al., 2022) menggunakan *Open Knowledge Maps* untuk menemukan dan mengonfirmasi tema utama dalam penelitian terkait pariwisata agro, agri, dan rural. (Handayani et al., 2022) juga menggunakan *Open Knowledge Maps* untuk memetakan perkembangan pengaplikasian kecerdasan buatan dalam *internet of things*. Selain itu, *Open Knowledge Maps* juga telah digunakan dalam berbagai bidang lainnya, seperti bisnis (Abbas & Mahsen Abdulrahman, 2023; Angkawijaya, 2022; Rantung et al., 2022; Wai, 2022), pemasaran (Francesca Chandra & Nadjib, 2023), teknik sipil (Seng Hansen, 2022), dan kesehatan (Delvechia Agustin et al., 2022; Farra Fazria & Dhamayanti, 2021; Haseeb Zia et al., 2021), untuk tujuan yang serupa.

Dengan membandingkan luaran penelitian dengan luaran *Open Knowledge Maps*, sebuah perangkat lunak yang sudah teruji dan dipakai oleh peneliti lainnya, luaran penelitian dapat dievaluasi keefektifan dan keakuratan pendekatan yang diusulkan, serta dapat diidentifikasi area dimana peningkatan dalam perangkat lunak dapat dilakukan.

3.9 Deployment Perangkat Lunak

Perangkat lunak yang merupakan hasil akhir dari penelitian ini *dideploy* dan berbasis web. Gambar 3.6 menunjukkan arsitektur dari perangkat lunak yang dikembangkan. Perangkat lunak terdiri dari tiga bagian utama, yaitu pengembangan layanan basis data, *backend*, dan *frontend*. Perangkat lunak ini juga berbasis web, karena memberikan kemudahan pada implementasi dan akses untuk masyarakat umum, agar perangkat apapun yang terhubung dengan internet dapat mengakses dan menggunakan perangkat lunak.



Gambar 3.6 Arsitektur perangkat lunak.

Layanan *backend* pada perangkat lunak merupakan sebuah API yang dibangun menggunakan bahasa pemrograman Python dan pustaka FastAPI. Adapun lingkup tugas dari API ini untuk mengambil data dari basis data akademis sesuai dengan kueri pencarian yang diberikan, melakukan pemodelan topik pada data yang telah diperoleh, dan menyimpan riwayat pencarian di basis data MySQL yang digunakan.

Lalu, untuk layanan *frontend* merupakan sebuah website yang dikembangkan dengan Next.js yang dihubungkan dengan API untuk menjadi antarmuka dalam melakukan pencarian *state-of-the-art* pada topik yang ditentukan oleh pengguna. Selain itu, layanan antarmuka ini juga yang merubah data yang diterima dari API setelah pemodelan dilakukan dalam bentuk visualisasi *nested bubble chart* dan *line chart*. Layanan *frontend* ini dihosting pada *cloud platform* Vercel, dan backend serta database dihosting pada server dengan sistem operasi Ubuntu di DigitalOcean dengan spesifikasi memori *random access memory* (RAM) sebesar 8GB dan *central processing unit* (CPU) dengan 4 *virtual core*. Perangkat lunak ini dapat diakses melalui tautan <https://ldaviewer.vercel.app/>.

BAB 4

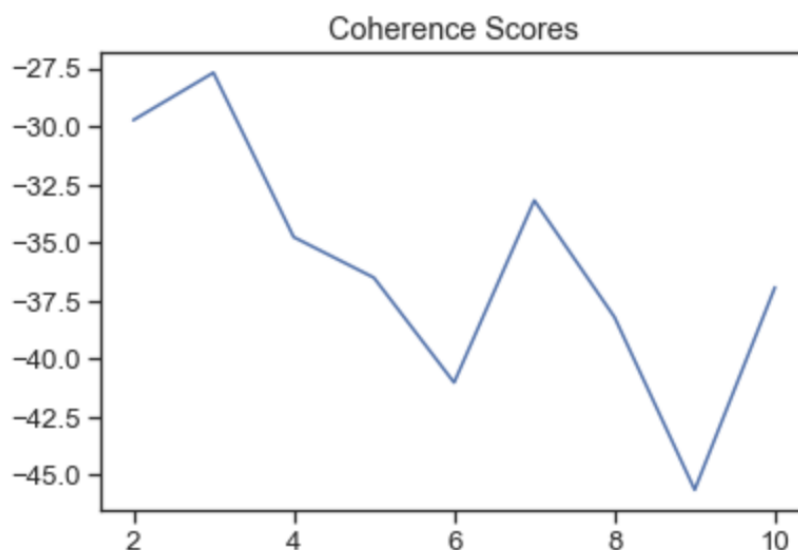
Hasil dan Pembahasan

4.1 Penentuan Algoritma

Untuk menentukan algoritma pemodelan topik yang akan digunakan, dua algoritma pemodelan topik standar, yakni LDA dan LSA akan diimplementasikan dan dibandingkan. Detil perbandingan antar kedua model dijabarkan pada sub-bab selanjutnya.

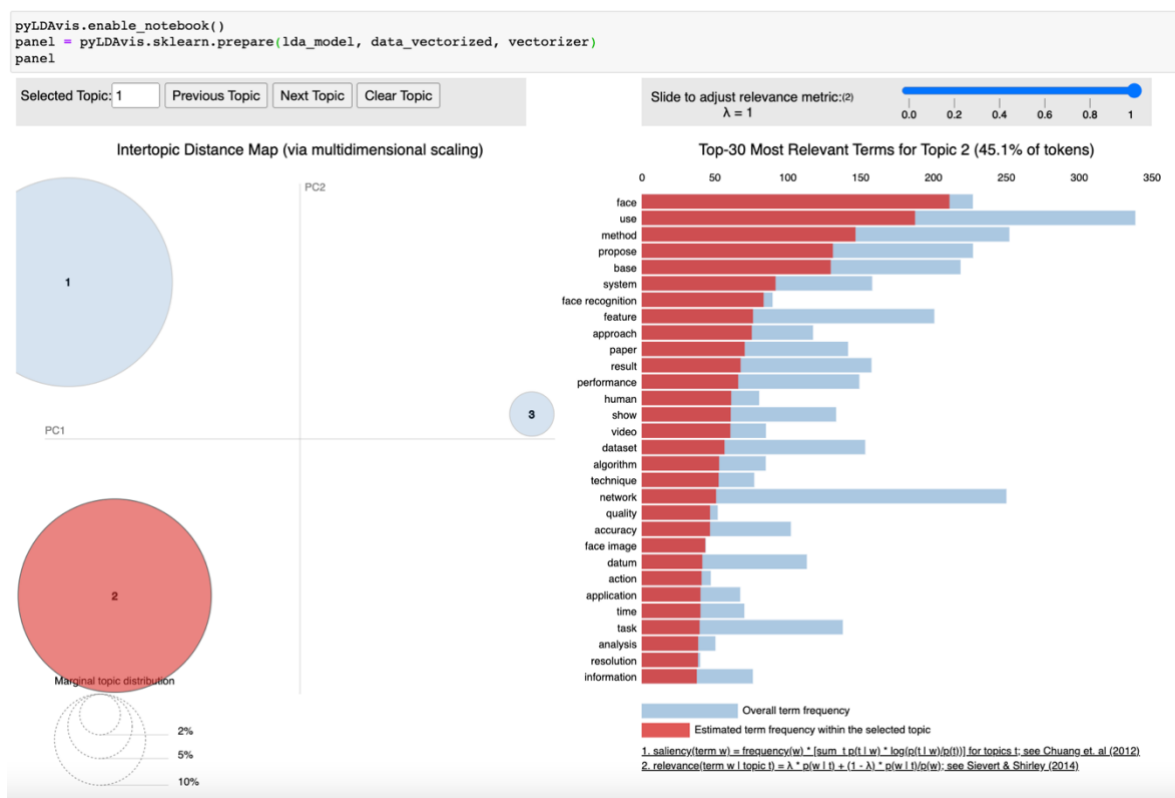
4.1.1 LDA

Penentuan nilai parameter α dan β pada model LDA dilakukan secara otomatis dengan metode *variational Bayes inference*. Lalu, nilai parameter K (jumlah topik) ditentukan menggunakan metode *elbow cut-off*, dimana nilai parameter K dengan skor koherensi *intrinsic UMass* tertinggi digunakan. Gambar 4.1 menunjukkan grafik yang memperlihatkan nilai koherensi untuk berbagai nilai K .



Gambar 4.1 Nilai koherensi model LDA untuk berbagai nilai K .

Dapat dilihat bahwa untuk pemodelan LDA, nilai koherensi terbaik diperoleh dengan $K = 3$, dengan nilai koherensi sebesar -27.69. Maka, nilai K ditetapkan menjadi 3 dan parameter tersebut digunakan untuk membangun model LDA. Memvisualisasikan luaran model dengan pustaka *pyLDavis* dalam bentuk *intertopic distance map*, dapat dilihat bahwa setiap topik yang dihasilkan oleh LDA tersegregasi dengan baik (lihat Gambar 4.2).



Gambar 4.2 Peta intertopik distance model LDA.

Hasil pemodelan LDA yang mengembalikan topik lalu dianalisis lebih lanjut menggunakan ekstraksi frasa yang menggunakan pola POS dan model BERT. Dari proses analisis lanjutan tersebut, dapat dilihat bahwa setiap topik berbicara tentang hal yang berbeda dan koheren.

Topic 0 Keyphrase with Weights [('recognition text', 0.5156), ('neural network result', 0.4976), ('network model', 0.4157), ('representation accuracy', 0.4025), ('system image', 0.3448)]
Topic 1 Keyphrase with Weights [('face recognition', 0.596), ('face use', 0.4623), ('dataset algorithm', 0.3361), ('feature approach', 0.3279), ('human show video', 0.323)]
Topic 2 Keyphrase with Weights [('fingerprint image', 0.6441), ('fingerprint recognition', 0.6334), ('biometric system segment', 0.6321), ('fingerprint minutia', 0.6221), ('recognition system', 0.5366)]

Gambar 4.3 Frasa dari setiap topik yang dihasilkan oleh model LDA.

Semisal, dapat dilihat dari luaran diatas bahwa terdapat pembicaraan mengenai pengenalan teks (topik pertama), pengenalan wajah (topik kedua), dan pengenalan sidik jari (topik ketiga) dalam bidang *image recognition*. Terdapat beberapa frasa-frasa yang masih “acak” atau tidak merepresentasikan suatu konsep asli, seperti *face use* atau *human show video*, yang lalu difiltrasi kembali dengan melakukan *entity linking* dengan KB ConceptNet dan Wikipedia. Hasil dari filtrasi dan normalisasi frasa dapat dilihat pada Tabel 4.1. Tabel 4.1 menunjukkan frasa yang diekstraksi sebelum difiltrasi dan dinormalisasi menggunakan *entity linking*, frasa yang telah difiltrasi dan dinormalisasi menggunakan *entity linking*

dengan basis data pengetahuan ConceptNet, dan frasa yang telah difiltrasi dan dinormalisasi menggunakan *entity linking* dengan basis data pengetahuan Wikipedia.

Tabel 4.1 Frasa Representatif untuk Setiap Topik LDA sebelum dan sesudah Filtrasi dan Normalisasi.

Topik	Frasa Representatif sebelum Filtrasi dan Normalisasi	Frasa Representatif setelah Filtrasi dan Normalisasi ConceptNet	Frasa Representatif setelah Filtrasi dan Normalisasi Wikipedia
1	<i>recognition text, neural network result, network model, representation accuracy, system image</i>	<i>neural network</i>	<i>Text recognition, Neural network, Network model, System image</i>
2	<i>face recognition, face use, dataset algorithm, feature approach, human show video</i>	<i>show video, face recognition</i>	<i>Face recognition, human video</i>
3	<i>fingerprint image, fingerprint recognition, biometric system segment, fingerprint minutia, recognition system</i>	<i>fingerprint recognition</i>	<i>Fingerprint recognition</i>

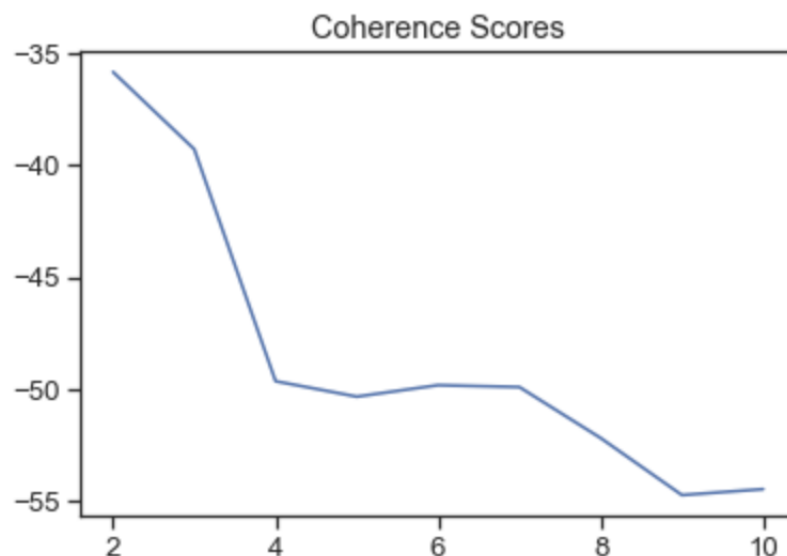
Untuk topik pertama, dapat dilihat bahwa filtrasi dengan ConceptNet hanya mengeluarkan frasa *neural network* sebagai satu-satunya frasa yang valid. Di sisi lain, dengan filtrasi Wikipedia, beberapa frasa seperti *text recognition*, *neural network*, *network model*, dan *system image* diperoleh. Untuk kedua topik lainnya, hasil filtrasi dan normalisasi dari kedua metode cukup mirip, dengan topik kedua hanya mengandung kata *face recognition*, *show video* (untuk ConceptNet), dan *human video* (untuk Wikipedia), dan topik ketiga hanya direpresentasikan melalui frasa *fingerprint recognition*.

Dengan proses filtrasi yang dilakukan, tema pembicaraan dari setiap topik semakin jelas dan terinci. Selain setiap frasa dari topik dapat dipastikan merupakan konsep nyata, proses ini juga menunjukkan potensi untuk pelabelan topik secara otomatis, dengan menunjukkan frasa-frasa representatif yang sudah difiltrasi dan dinormalisasi. Mengevaluasi

luaran topik serta frasa representatif untuk setiap topik, dapat disimpulkan bahwa LDA mampu menemukan topik-topik dari abstrak-abstrak penelitian dengan baik, dengan artian setiap topik tersegregasi dengan baik.

4.1.2 LSA

Selain LDA, LSA juga diimplementasikan agar dapat menjadi dasar penentuan algoritma pemodelan topik yang akan digunakan untuk luaran penelitian. Dalam implementasinya, LSA hanya membutuhkan penentuan parameter K , atau jumlah topik. Nilai parameter K ditentukan dengan cara yang serupa dengan LDA, yakni menggunakan metode *elbow cut-off*. Gambar 4.4 menunjukkan grafik yang memperlihatkan nilai koherensi untuk berbagai nilai K .



Gambar 4.4 Nilai koherensi model LSA untuk berbagai nilai K .

Untuk pemodelan LSA, nilai koherensi terbaik diperoleh dengan $K = 2$, dengan nilai koherensi sebesar -35.80. Maka, nilai K akan ditetapkan menjadi 2. Setelah pemodelan dilakukan, mirip dengan LDA, ekstraksi frasa menggunakan model BERT dilakukan untuk membantu dalam proses pelabelan topik. Adapun luaran dari frasa yang diekstraksi dapat dilihat pada Gambar 4.5.

```
Topic 0 Keyphrase with Weights [('face recognition', 0.5384), ('face method', 0.4644), ('feature base', 0.3885), ('neural network', 0.3693), ('performance data set', 0.3686)]
Topic 1 Keyphrase with Weights [('recognition system', 0.4808), ('recognition face', 0.4792), ('recognition method', 0.4586), ('image quality', 0.4048), ('image resolution', 0.3721)]
```

Gambar 4.5 Frasa dari setiap topik yang dihasilkan oleh model LSA.

Filtrasi dan normalisasi frasa yang dihasilkan oleh model LSA juga dilakukan menggunakan ConceptNet dan Wikipedia. Tabel 4.2. menunjukkan frasa sebelum difiltrasi dan dinormalisasi, dan luaran filtrasi dan normalisasi frasa melalui basis data ConceptNet dan Wikipedia.

Tabel 4.2 Frasa Representatif untuk Setiap Topik LSA sebelum dan sesudah Filtrasi dan Normalisasi.

Topik	Frasa Representatif sebelum Filtrasi dan Normalisasi	Frasa Representatif setelah Filtrasi dan Normalisasi ConceptNet	Frasa Representatif setelah Filtrasi dan Normalisasi Wikipedia
1	<i>face recognition, face method, feature base, neural network, performance set</i>	<i>neural network, face recognition</i>	<i>Neural network, Face recognition</i>
2	<i>recognition system, recognition face, recognition method, image quality, image resolution</i>	<i>face recognition, image resolution, image quality</i>	<i>Face recognition, Image resolution, Image quality</i>

Dapat dilihat bahwa untuk kedua topik, LSA tampak kesulitan untuk mensegregasikan kedua topik, ditunjukkan dari frasa-frasa yang mirip antar kedua topik yang sama-sama berbicara tentang pengenalan wajah. Melakukan filtrasi dan normalisasi lebih lanjut dengan ConceptNet dan Wikipedia, dapat dikonfirmasi bahwa kedua topik tidak terlalu tersegregasi dengan baik, dikarenakan kedua topik dapat disimpulkan berbicara mengenai pengenalan wajah.

4.1.3 Perbandingan Model LDA dan LSA

Dari hasil pemodelan, Tabel 4.3 menunjukkan tabel perbandingan hasil antara kedua algoritma yang digunakan. Adapun hasil yang dipaparkan pada tabel berupa algoritma yang diuji, yaitu LDA dan LSA, nilai koherensi yang diperoleh dari masing-masing model, dan apakah luaran topik tersegregasi dengan baik.

Tabel 4.3 Perbandingan Hasil antara Model LDA dan LSA.

Algoritma	Nilai Koherensi	Segregasi Topik yang Baik
Latent-Dirichlet Allocation	-27.69	Ya
Latent Semantic Analysis	-35.80	Tidak

Dari performa kedua model dalam menemukan topik dalam data abstrak yang diberikan, dapat dilihat bahwa LDA mengungguli LSA di beberapa aspek, seperti nilai koherensi dan pemisahan topik. Model LDA yang digunakan memiliki nilai koherensi sebesar -27.69, sedangkan nilai koherensi dari model LSA lebih kecil, yakni -35.80. Selain itu, untuk luaran topik, LDA menghasilkan tiga topik, masing-masing merepresentasikan topik yang berbeda, yaitu mengenai jaringan saraf tiruan, pengenalan wajah, dan pengenalan sidik jari.

Di sisi lain, LSA menghasilkan dua topik, tetapi, kedua topik yang dihasilkan merepresentasikan hal yang sama, yakni pengenalan wajah, sehingga pemisahan topik dari model LSA kurang sempurna. Walaupun, perlu dicatat bahwa untuk proses ekstraksi frasa yang dapat digunakan sebagai pelabelan topik, kedua model dapat menghasilkan frasa yang representatif. Mempertimbangkan hasil yang diperoleh dari eksperimen awal dan tinjauan literatur yang telah dilakukan yang menunjukkan prevalensi LDA dalam riset yang serupa, LDA akan digunakan sebagai algoritma utama untuk melakukan pemodelan topik dalam penelitian ini.

4.2 Hasil Implementasi Perangkat Lunak

Dengan terpilihnya algoritma yang akan dipakai, serta pipeline pengumpulan data, metode pemodelan dan visualisasi yang telah dipaparkan pada sub-bab sebelumnya, sebuah antarmuka pencarian yang dapat dipakai untuk mengetahui topik-topik terkini dari topik penelitian yang ingin dicari dihasilkan dan dapat diakses di <https://ldaviewer.vercel.app/>. Adapun fitur-fitur utama dari perangkat lunak dijelaskan pada sub-bab berikut.

4.2.1 Pencarian Standar

Perangkat lunak menyediakan fitur “pencarian standar” untuk pengguna, dimana pengguna dapat langsung memasukkan kata kunci yang ingin dicari dan menekan tombol pencarian. Gambar 4.6 menunjukkan tangkapan layar dari fitur pencarian standar.

Welcome to LDASViewer!

Get started by searching your topic of interest!

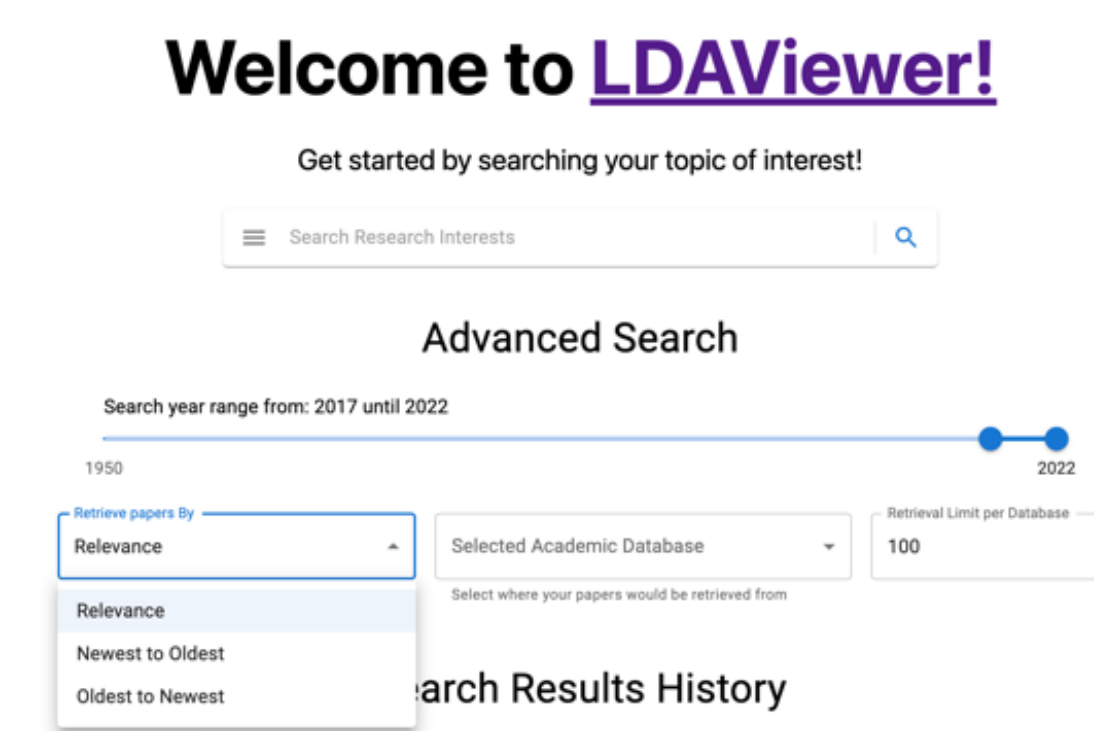


Gambar 4.6 Fitur pencarian standar.

Pada fitur ini, pengguna dapat memasukkan kueri pencarian tentang topik apapun pada *search bar* yang disediakan. Lalu, ketika pengguna menekan tombol pencarian yang berbentuk kaca pembesar, atau menekan tombol *enter* pada *keyboard*, maka pencarian akan dilakukan. Pencarian akan melakukan proses pengambilan data berdasarkan kueri pengguna dari semua basis data, yakni CORE, arXiv, Emerald, ScienceOpen, dan Garuda dengan batasan abstrak artikel yang dikumpulkan per basis data sebesar 100 dan pengambilan yang disortir berdasarkan relevansi artikel. Data yang telah dikumpulkan akan digunakan untuk melakukan pemodelan topik dan ekstraksi label dari luaran pemodelan topik secara otomatis.

4.2.2 Pencarian *Advanced*

Fitur kedua yang disediakan perangkat lunak adalah fitur “pencarian *advanced*”. Fitur ini menyediakan pengguna dengan kontrol yang lebih banyak untuk mengatur parameter pencarian *state-of-the-art* pada topik yang diinginkan. Gambar 4.7 dan 4.8 merupakan tangkapan layar yang menunjukkan fitur pencarian *advanced* pada perangkat lunak.



Gambar 4.7 Fitur pencarian *advanced* (a).

Pengguna dapat mengakses fitur pencarian ini dengan menekan tombol dengan ikon *hamburger menu* di sebelah kiri *search bar* yang disediakan. Pada fitur ini, terdapat beberapa parameter pencarian tambahan yang dapat diatur oleh pengguna. Pertama, terdapat sebuah *slider* rentang tahun yang dapat diatur agar pencarian artikel dilakukan hanya pada rentang tahun tertentu. Hal ini berguna jika pengguna ingin menganalisis penelitian pada topik dalam rentang waktu tertentu, seperti hanya pada lima tahun terakhir untuk melihat pembicaraan terkini, atau rentang tahun dimana sebuah fenomena terjadi untuk melihat perubahan paradigma penelitian karena peristiwa tertentu. Selain itu, pengguna juga dapat memilih untuk memperoleh konten artikel berdasarkan *relevance* (relevansi), *newest to oldest* (tanggal publikasi yang diurutkan dari terbaru hingga tertua), dan *oldest to newest* (tanggal publikasi yang diurutkan dari tertua hingga terbaru) masing-masing artikel.

Welcome to LDAVIEWER!

Get started by searching your topic of interest!

Search Research Interests

Advanced Search

Search year range from: 2017 until 2022

1950 2022

Retrieve papers By: Relevance

Select how you want your papers to be retrieved

Selected Academic Database

CORE
arXiv
Emerald
ScienceOpen

Retrieval Limit per Database: 100

ID	Action	Action
----	--------	--------

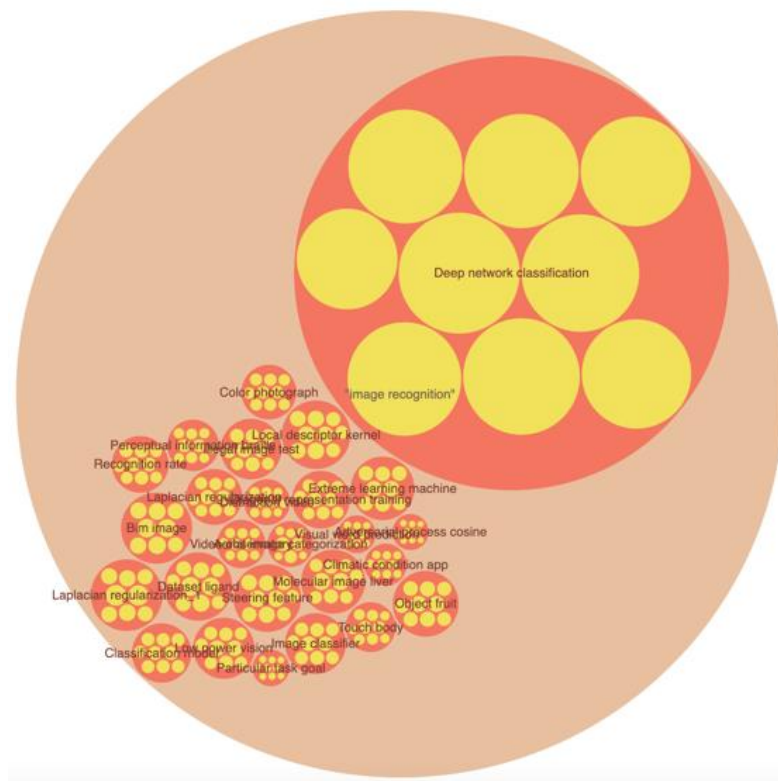
Gambar 4.8 Fitur pencarian *advanced* (b).

Pengguna juga dapat memilih basis data akademis yang ingin digunakan untuk melakukan pencarian. Disini, pilihan basis data akademis yang disediakan adalah CORE, arXiv, Emerald, ScienceOpen, dan Garuda. Fitur ini merupakan hal yang penting karena beberapa basis data mungkin kurang cocok atau tidak menaruh fokus pada topik penelitian yang ingin diselidiki, sehingga pengguna dapat memilih basis data akademis yang sesuai. Selain itu, pengguna juga dapat menentukan batasan jumlah pengambilan artikel pada setiap basis data. Mirip dengan pencarian standar, setelah pengguna sudah selesai mengkonfigurasi parameter pencarian, perangkat lunak akan melakukan proses pengambilan data sesuai dengan parameter yang telah ditentukan dan melakukan pemodelan topik serta ekstraksi label dan menghasilkan visualisasi secara otomatis dengan data yang telah dikumpulkan.

4.2.3 Visualisasi *Nested Bubble Chart*

Setelah melakukan pengambilan data, pemodelan topik, dan ekstraksi label, perangkat lunak memvisualisasikan luaran pemodelan topik dan ekstraksi label dalam bentuk visualisasi *nested bubble chart* seperti yang ditunjukkan pada Gambar 4.9.

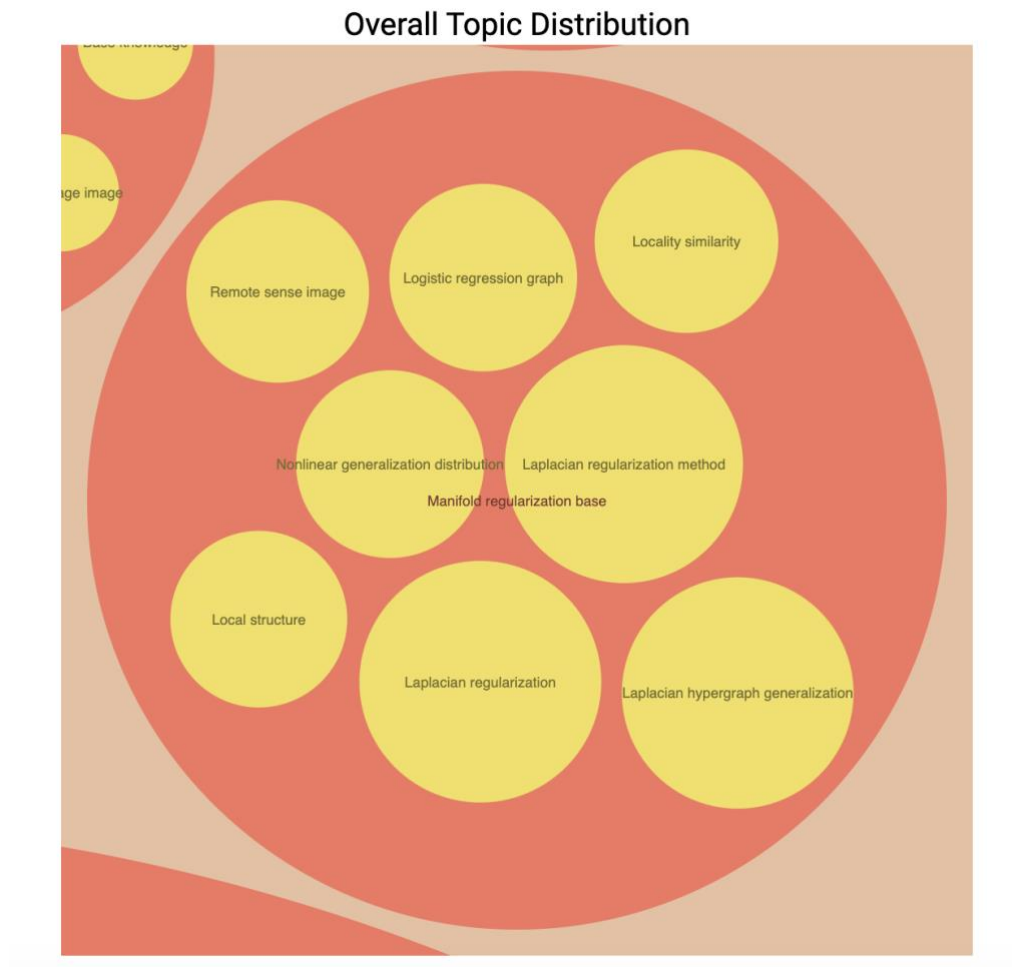
LDA results for "image recognition"



Gambar 4.9 Sampel luaran visualisasi *nested bubble chart*.

Visualisasi ini menunjukkan distribusi jumlah dokumen untuk setiap topik. Pada visualisasi ini, sebuah *bubble* berwarna merah merepresentasikan sebuah topik, dan ukuran sebuah *bubble* merupakan jumlah dokumen yang dikategorikan sebagai topik tersebut. Selain itu, terdapat *bubble* berwarna kuning di dalam setiap *bubble* yang merupakan frasa-frasa lainnya yang merepresentasikan topik tertentu (lihat Gambar 4.10). Jika sebuah topik ditekan, maka rincian topik tersebut akan dimunculkan melalui *bubble* berwarna kuning tersebut.

LDA results for "image recognition"

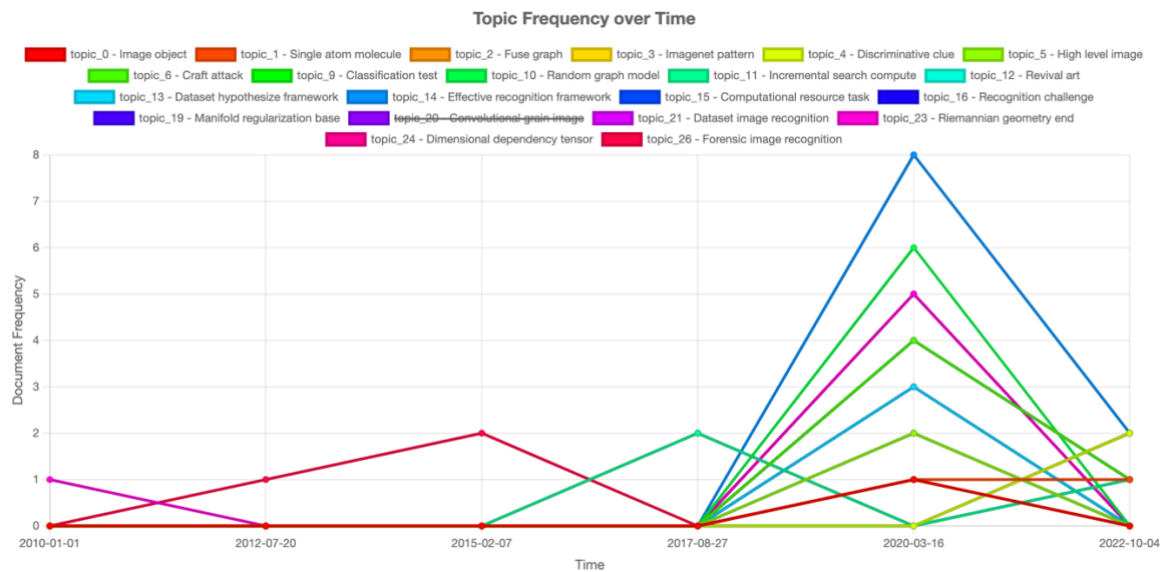


Gambar 4.10 Sampel luaran rincian visualisasi *nested bubble chart*.

Bubble rincian topik berwarna kuning ini merupakan frasa-frasa lainnya yang dapat merepresentasikan topik. Sebagaimana LDA pada dasarnya mengasumsikan sebuah topik adalah percampuran kata-kata yang saling terkait pada korpus, visualisasi ini memberikan informasi rinci tentang sebuah topik dengan campuran frasa-frasa yang saling terkait. Diharapkan dengan visualisasi ini, dapat menunjukkan *gap* atau kesenjangan dalam literatur, semisal, dimana terdapat kurangnya penelitian dalam topik tertentu.

4.2.4 Visualisasi *Line Chart*

Selain memvisualisasikan luaran pemodelan topik dan ekstraksi label dalam bentuk visualisasi *nested bubble chart*, visualisasi *line chart* juga dihasilkan dari perangkat lunak (lihat Gambar 4.11).



Gambar 4.11 Sampel luaran visualisasi *line chart*.

Visualisasi ini menunjukkan tren distribusi jumlah dokumen untuk setiap topik dari waktu-ke-waktu. Pada visualisasi ini, axis vertikal pada grafik menunjukkan frekuensi topik dan axis horizontal menunjukkan waktu. Setiap garis pada grafik merepresentasikan sebuah topik, dan tren atau perubahan datanya dari tahun ke tahun dapat dilihat pada perubahan nilai di titik-titik tertentu pada setiap garis. Visualisasi ini dapat menunjukkan tren pembicaraan dari setiap topik, dan juga berpotensi untuk menunjukkan topik-topik yang sedang memiliki kenaikan atau penurunan drastis dalam tren riset, sehingga dapat menunjukkan perubahan paradigma pada bidang riset tertentu. Selain itu, dengan visualisasi ini, diharapkan dapat menunjukkan arah riset saat ini dan prediksi untuk arah riset selanjutnya.

4.2.5 Riwayat Pencarian

Fitur keempat pada perangkat lunak adalah riwayat pencarian (lihat Gambar 4.12). Riwayat pencarian yang pernah dilakukan oleh pengguna perangkat lunak disajikan dalam bentuk sebuah tabel yang mengandung *ID*, yang merupakan sebuah angka untuk mengidentifikasi riwayat pencarian secara unik, *search query* yakni kueri pencarian yang dilakukan, *created at* yaitu waktu pencarian dilakukan, dan *action* yang merupakan aksi yang dapat dilakukan oleh pengguna pada riwayat pencariannya.

Search Results History

ID	Search Query	Created At	Action
5	"exoplanet"	2022-09-04T00:19:27	VIEW
6	"competitive advantage"	2022-09-06T09:06:02	VIEW
7	"image recognition"	2022-09-20T09:15:21	VIEW
8	"image recognition"	2022-09-20T09:51:41	VIEW

Gambar 4.12 Sampel riwayat pencarian.

Pada saat sebuah pencarian dilakukan oleh seorang pengguna, hasil pencarian akan disimpan dalam sebuah basis data *MySQL*. Data yang disimpan ini diperoleh dan ditunjukkan pada riwayat pencarian. Lalu, jika pengguna menekan tombol “View” pada salah satu pencarian sebelumnya, visualisasi *nested bubble* dan *line chart* pemodelan topik yang ditunjukkan sebelumnya akan dimunculkan oleh perangkat lunak. Riwayat pencarian merupakan fitur yang penting bagi pengguna karena mereka dapat dengan mudah mengakses informasi yang telah dicari sebelumnya, menghemat waktu dan usaha. Selain itu, fitur ini juga dapat membantu pengguna untuk meninjau riset yang dilakukan sebelumnya dari waktu ke waktu.

4.2.6 Detil Pencarian dan Pemodelan

Gambar 4.13 menunjukkan fitur terakhir yang disediakan oleh perangkat lunak adalah ditunjukkannya detil pencarian dan pemodelan yang telah dilakukan. Pada gambar, dapat dilihat bahwa perangkat lunak memperlihatkan jumlah dokumen yang diproses dalam pemodelan, perinician jumlah dokumen yang digunakan dari setiap basis data yang digunakan, nilai koherensi dari model pemodelan topik yang digunakan, dan waktu yang dibutuhkan untuk melakukan proses pencarian data hingga mengeluarkan luaran akhir pemodelan topik. Selain itu, juga tersedia tombol untuk mengunduh dokumen-dokumen yang diproses untuk kueri pencarian pengguna dalam bentuk dokumen *comma-separated values* (CSV) yang sampel kontennya ditunjukkan pada Tabel 4.4.

Total Documents Processed: 391

Document Details

Database: arXiv 198 documents

Database: CORE 193 documents

Number of Topics: 25

Coherence Score: -33.36

Process Time: 5.91 minutes

DOWNLOAD PROCESSED DOCUMENTS IN CSV

Gambar 4.13 Sampel detail pencarian dan pemodelan.

Tabel 4.4 Sampel Konten Dokumen CSV dari Perangkat Lunak.

<i>Title</i>	<i>Source</i>	<i>Abstract</i>	<i>Assigned Topic</i>	<i>Published Date</i>	<i>Main Topic</i>	<i>Sub Keyphrases</i>
<i>Entanglement fidelity and measure of entanglement</i>	arXiv	<i>The notion of entanglement fidelity is to measure...</i>	7	2019-11-25 11:58:49	<i>Quantum state information</i>	<i>Pure state entanglement, State entanglement, Entanglement phase, ...</i>

Tabel 4.4 menyajikan sampel konten dokumen CSV yang dihasilkan oleh perangkat lunak untuk kueri pencarian “*quantum entanglement*”. Dokumen CSV tersebut menyediakan beberapa informasi tentang dokumen, di antaranya judul artikel (*title*), sumber basis data dimana artikel diperoleh (*source*), dan abstrak dari artikel (*abstract*). Selain itu, hasil pemodelan topik dari masing-masing artikel juga ditampilkan melalui informasi topik yang ditugaskan (*assigned topic*), tanggal publikasi (*published date*), topik utama (*main topic*) dan frasa kunci lainnya yang dapat digunakan untuk merepresentasikan topik (*sub keyphrases*). Dalam tabel ini hanya ditampilkan sebagian dari abstrak dan sub frasa menggunakan tanda elipsis.

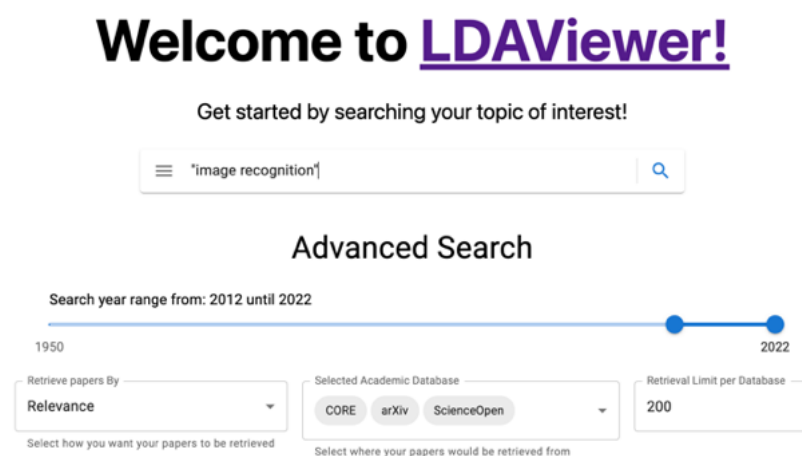
Melihat detail dari proses pencarian dan pemodelan yang telah dilakukan merupakan hal penting karena hal ini dapat memberikan informasi yang diperlukan untuk mengevaluasi kualitas dari hasil yang diperoleh. Informasi seperti jumlah dokumen yang diproses dari pemodelan, nilai koherensi dari model pemodelan topik, dan waktu yang dibutuhkan untuk melakukan proses pencarian data dapat digunakan untuk mengevaluasi efektivitas dari metode yang digunakan dalam proses pemodelan. Dengan tombol untuk mengunduh dokumen-dokumen yang diproses dalam bentuk dokumen CSV juga dapat memudahkan

pengguna dalam meneliti atau mengevaluasi hasil yang diperoleh, serta memudahkan dalam melakukan analisis lanjutan. Fitur ini juga ditujukan agar dapat membuat hasil pemodelan yang dilakukan lebih transparan, agar penelitian yang menggunakan perangkat lunak ini dapat direproduksi dan divalidasi oleh peneliti-peneliti lainnya sebaik mungkin.

4.3 Hasil Evaluasi Perangkat Lunak

4.3.1 Hasil Pengujian Awal

Untuk memastikan bahwa perangkat lunak yang dibuat dapat menghasilkan topik yang koheren, perangkat lunak diuji menggunakan kata kunci pencarian yang sama pada sub-bab 4.1.2, yakni “*image recognition*”. Selain itu, basis data yang digunakan untuk pengujian ini adalah CORE, arXiv, dan ScienceOpen. Pengumpulan data abstrak dibatasi pada 200 artikel dari setiap basis data yang dipublikasikan pada rentang 10 tahun kebelakang (2012-2022). Parameter-parameter pencarian yang telah ditentukan untuk pencarian dapat dilihat pada Gambar 4.14.



The screenshot displays the LDAViewer web interface. At the top, it says "Welcome to **LDAViewer!**" followed by the instruction "Get started by searching your topic of interest!". Below this is a search bar containing the text "image recognition" and a magnifying glass icon. Under the search bar, the heading "Advanced Search" is visible. Below the heading, there is a horizontal timeline slider for "Search year range from: 2012 until 2022", with markers at 1950 and 2022. Below the timeline, there are three main sections: "Retrieve papers By" with a dropdown menu set to "Relevance", "Selected Academic Database" with buttons for "CORE", "arXiv", and "ScienceOpen", and "Retrieval Limit per Database" with a text input set to "200". Small text below these sections reads "Select how you want your papers to be retrieved" and "Select where your papers would be retrieved from".

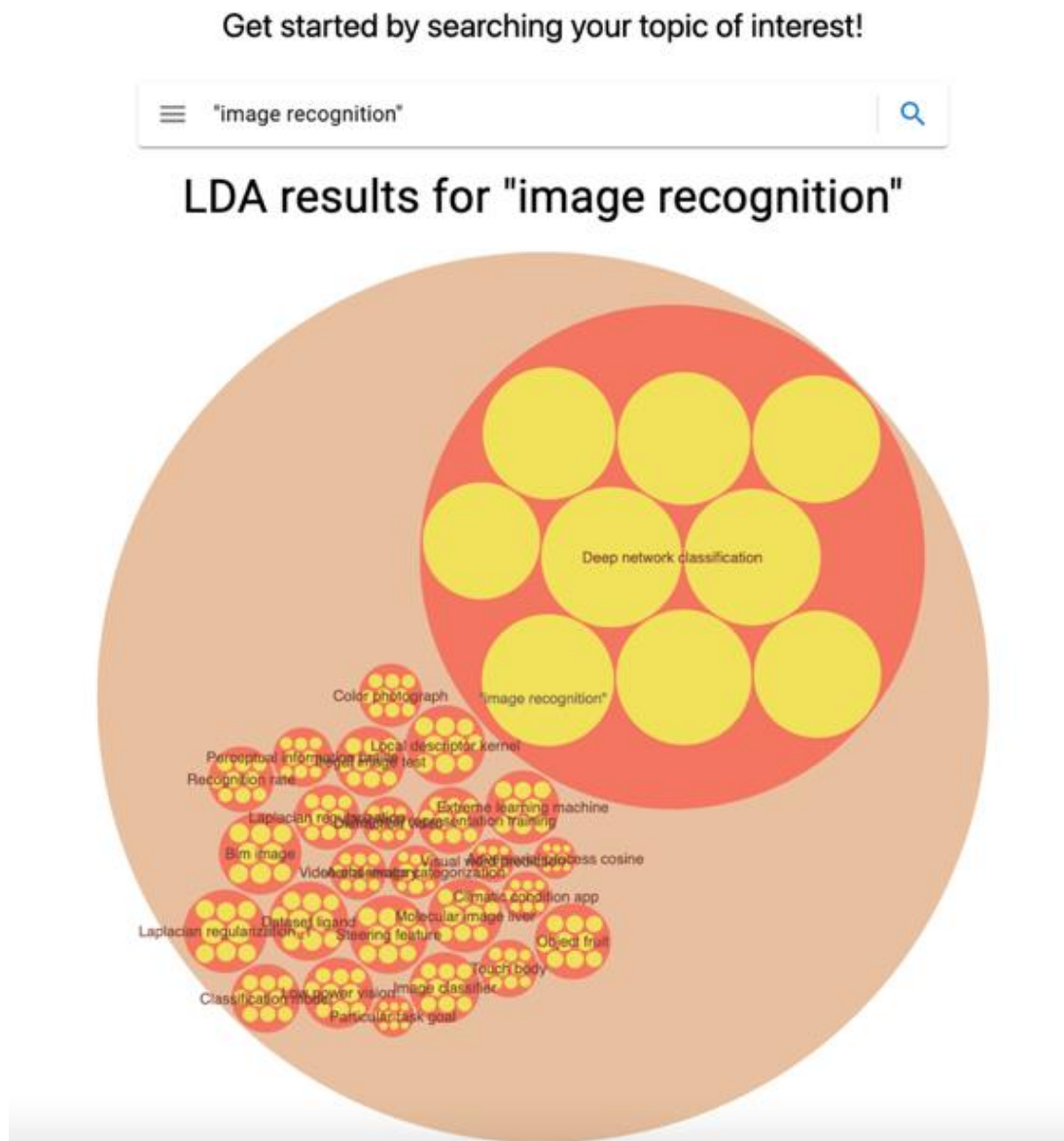
Gambar 4.14 Parameter yang digunakan untuk melakukan pencarian pada perangkat lunak.

Luaran dari hasil pencarian pada perangkat lunak diperoleh dalam waktu 14 menit, dengan total jumlah artikel yang diproses sebanyak 594 (200 dari ScienceOpen, 199 CORE, dan 195 dari arXiv). Nilai koherensi yang didapatkan adalah -19.07 (lihat Gambar 4.15).

Total Documents Processed: 594
Document Details
Database: ScienceOpen 200 docs
Database: CORE 199 docs
Database: arXiv 195 docs
Coherence Score: -19.07
Process Time: 14.29 minutes

Gambar 4.15 Tangkapan layar detail parameter pemodelan dari perangkat lunak.

Visualisasi pemodelan topik dari data yang telah dikumpulkan dapat dilihat pada Gambar 4.16 dan 4.17. Gambar 4.16 merupakan visualisasi *nested bubble chart* yang menunjukkan topik-topik yang ditemukan dari perangkat lunak serta frekuensinya.



Gambar 4.16 Visualisasi *nested bubble chart* dari hasil pemodelan topik dalam bidang pengenalan gambar.

Terdapat 26 topik yang dihasilkan, dengan penjelasan masing-masing topik dijabarkan pada Tabel 4.5. Tabel 4.5 menunjukkan luaran dari perangkat lunak pengenalan gambar yang dikelompokkan berdasarkan topik, label, dan frasa. Topik adalah nomor yang digunakan untuk mengidentifikasi topik, label adalah frasa utama yang digunakan untuk merepresentasikan topik, dan frasa adalah frasa-frasa lainnya yang dapat digunakan untuk merepreentasikan topik. Contoh dari topik yang dibahas dalam tabel 4.4 adalah topik pertama, yang direpresentasikan oleh label "*Deep network classification*" dan frasa-frasa seperti "*Deep network classification*", "*image classification*", "*image recognition*", "*model recognition*", dan "*neural network*".

Tabel 4.5 Luaran Topik Perangkat Lunak untuk Bidang Pengenalan Gambar.

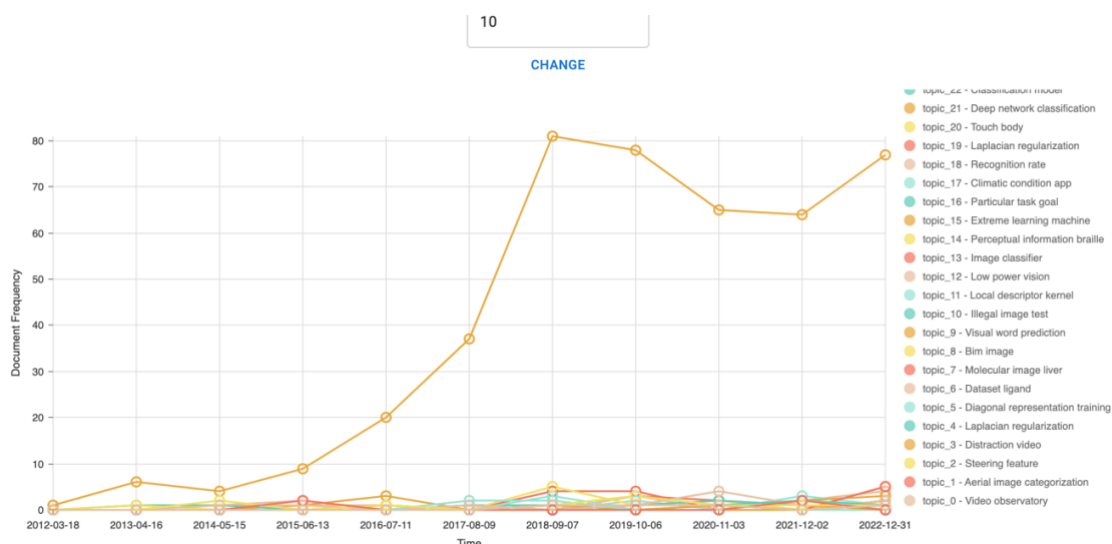
Topik	Label	Frasa
1	<i>Deep network classification</i>	<i>Deep network classification, image classification, image recognition, model recognition, neural network</i>
2	Pengenalan buah	<i>Object fruit, tomato fruit, size fruit, tomato color</i>
3	Pengenalan gambar dengan gambar berwarna	<i>Color photogrpah, color image, lighting condition, different lighting condition, photograph assist</i>
4	<i>Kernel</i>	<i>Local descriptor kernel, local descriptor, descriptor kernel, training representation, spd representation</i>
5	<i>Extreme machine learning</i>	<i>Extreme machine learning, accuracy mnist, effective hide neuron, random feature image</i>
6	<i>Adversarial attacks</i>	<i>Adversarial process cosine, adversarial process subsidiary, identification image, subsidiary discriminative</i>
7	Pengaplikasian <i>Image recognition</i> pada iklim	<i>Climate condition app, abundant rainfall, dense vegetation kind</i>
8	Pengenalan kata visual	<i>Visual word prediction, visual word ear, visual cue, conceptual region prediction</i>
9	<i>Diagonal representation training</i>	<i>Diagonal representation optimization, class representation, extra class representation</i>
10	Pengaplikasian <i>Image recognition</i> pada bidang forensik	<i>Illegal image test, forensic image recognition, law enforcement, enforcement agency</i>
11	Pengenalan <i>Braille</i>	<i>Perceptual information braille, braille dot, dot braille, perceptual information hiding</i>
12	Regularisasi <i>Lapalacian</i>	<i>Laplacian regularization, graph laplacian, multiple graph regularization</i>
13	Pengenalan distraksi untuk kendaraan otonom	<i>Distraction video, severity distraction, distraction severity, vehicle driver, autonomous vehicle driver, autonomous vehicle control, accident driver</i>

Topik	Label	Frasa
14	Pengenalan gambar udara	<i>Aerial image categorization, aerial image discovery, aerial image refine, aerial image sub</i>
15	Pengenalan gambar molekular	<i>Molecular image, molecular image model, molecular graph, sequential representation molecule, molecular structure</i>
16	Metodologi pengenalan gambar	<i>Softmax output softmax, constant input estimation, baseline experiment, outperform baselines, local model</i>
17	Metodologi pengklasifikasi gambar	<i>Image classifier, shot learning, transferable knowledge, dataset structure shoot, category diversity, dataset novel category</i>
18	Kerangka pengenalan gambar	<i>General reinforcement formulation, useful information order, search framework, abstract concept, policy framework</i>
19	Pengenalan gambar dengan sumber daya rendah	<i>Low power image, mobile phone recognition, good technology battery, low power, system drone</i>
20	Pengaplikasian pengenalan gambar pada pengemudian	<i>Steering device, feature recognition, training strategy</i>
21	Pengenalan video	<i>Video observatory, video imagery, automate recognition result, autonomous sensor study</i>
22	Pengenalan gambar terkait tanaman	<i>Dataset ligand, real time plant, principal component analysis, different view compound</i>
23	<i>Building information models (BIM)</i>	<i>Bim image, widespread attention bim, conventional bim image, computational architecture</i>
24	Performa model	<i>Recognition rate, efficiency accuracy, task memory, image functionality</i>

Topik	Label	Frasa
25	Regularisasi <i>Laplacian</i>	<i>Laplacian regularization, graph laplacian, multiple graph regularization</i>
26	Klasifikasi gambar	<i>Appriopriate level taxonomy, taxonomoical organization class, familiar class classification, chemical structure</i>

Dapat dilihat dari visualisasi yang dilakukan, topic *deep network classification* yang berbicara tentang metode utama yang digunakan dalam pengenalan gambar merupakan topik dengan artikel terbanyak dalam bidang ini. Hal ini masuk akal karena pengenalan gambar dilakukan menggunakan *deep neural networks*. Di sisi lain, dapat dilihat bahwa topik 6 tentang *adversarial attacks* memiliki *bubble* yang sangat kecil. Hal ini dapat menandakan bahwa belum banyak riset yang berbicara tentang hal tersebut. Ini juga dapat diartikan sebagai sebuah *gap* dalam literatur, dimana terdapat kurangnya penelitian dalam topik tersebut. Selain dari wawasan yang telah diperoleh sebelumnya, secara subjektif, dapat dilihat bahwa setiap topik dapat dilihat tersegregasi dan tidak menumpang tindih satu dengan yang lainnya.

Gambar 4.17 merupakan sebuah *line chart* yang menunjukkan tren frekuensi dokumen setiap topik untuk N contiguous sub-interval, dimana N adalah nilai yang dapat ditentukan oleh pengguna perangkat lunak. Dalam kasus ini, N adalah 10, sesuai dengan rentang waktu tahun pencarian (2012 - 2022).



Gambar 4.17 Visualisasi *line chart* dari hasil pemodelan topik dalam bidang pengenalan gambar.

Dapat dilihat bahwa terdapat tren kenaikan untuk topik *molecular image recognition* (topik 15 pada tabel) dari tahun ke tahun hingga saat ini, yang dapat menandakan bahwa penelitian untuk topik tersebut sedang tren. Dengan hasil di atas, dapat dilihat bahwa perangkat lunak dapat menghasilkan luaran topik *state-of-the-art* dalam sebuah topik penelitian, dan pengaplikasiannya pun dapat diuji coba pada topik-topik riset pada bidang ilmu apapun dikarenakan sifat pemrosesan yang *context independent*.

4.3.2 Hasil Survei Perangkat Lunak

Perangkat lunak yang dikembangkan diuji lebih lanjut untuk memastikan bahwa secara kualitatif layak digunakan dalam penelitian dan oleh masyarakat umum. Evaluasi kualitatif ini meliputi kuesioner yang dibuat untuk menilai dua dimensi utama dari hasil penelitian, yaitu aplikasi (perangkat lunak) dan hasil pemodelan topik. Kuesioner dibagikan dan telah diisi oleh 52 individu yang merupakan peneliti mahasiswa, dengan hasil survei yang dapat dilihat pada lampiran A. Adapun analisis dari hasil survei yang diperoleh akan dijelaskan selanjutnya. Terdapat 52 responden yang berasal dari berbagai fakultas dengan minat riset yang beragam (lihat Gambar 4.18).

Fakultas
52 responses

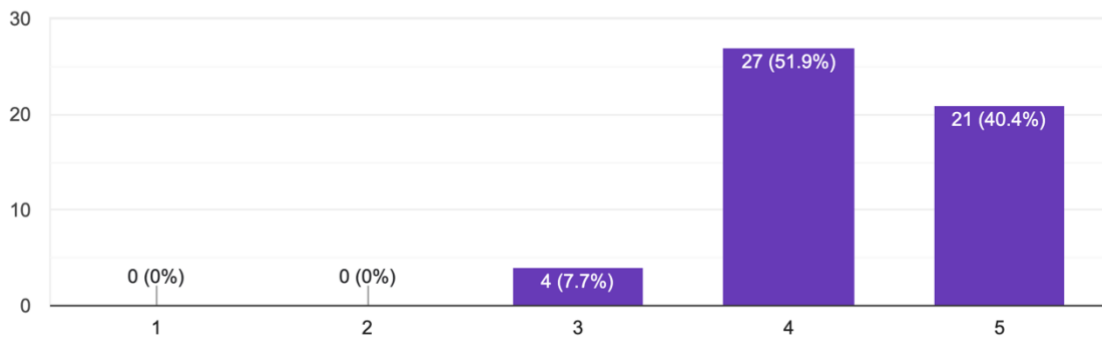


Gambar 4.18 Distribusi fakultas responden dalam survei evaluasi luaran penelitian.

Mayoritas responden merupakan peneliti dari fakultas teknologi industri, namun terdapat juga banyak responden dari fakultas lain seperti teknik sipil dan perencanaan, psikologi dan ilmu sosial budaya, matematika dan ilmu pengetahuan alam, kedokteran, ilmu agama islam, bisnis dan ekonomika, serta pendidikan.

Antarmuka aplikasi ini mudah untuk dinavigasi

52 responses

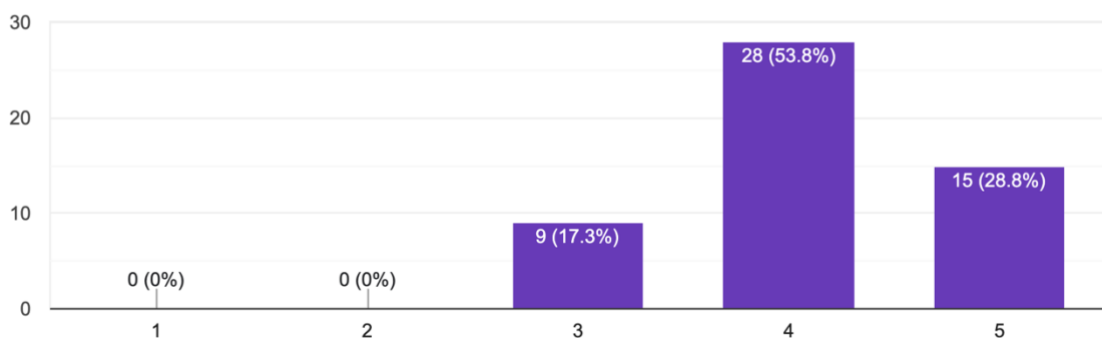


Gambar 4.19 Distribusi nilai kemudahan navigasi pada aplikasi dalam survei evaluasi luaran penelitian.

Mengenai dimensi aplikasi, khususnya dalam hal kemudahan navigasi, dapat dilihat pada Gambar 4.19 bahwa rata-rata skor yang diberikan adalah 4,3 dari skala 1-5, dengan 5 sebagai skor terbaik, dengan distribusi yang terbanyak pada skala keempat dan kelima. Skor ini mengimplikasikan bahwa aplikasi secara umum dapat dengan mudah digunakan dan dinavigasi. Hal ini dapat disebabkan oleh fakta bahwa perangkat lunak ini dirancang untuk memiliki antarmuka yang intuitif. Elemen-elemen antarmuka seperti *search bar* untuk pencarian digunakan karena akrab digunakan pada perangkat lunak populer lainnya seperti Google, serta *placeholder* dan label yang memberikan pengarahannya instruksi penggunaan bagi pengguna sehingga memudahkan pengguna dalam menavigasi aplikasi.

Saya bisa mendapatkan informasi dengan mudah dari antarmuka aplikasi ini

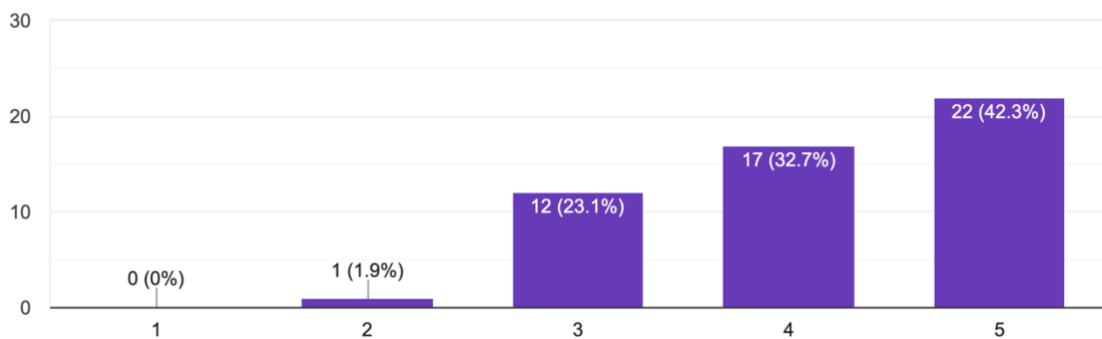
52 responses



Gambar 4.20 Distribusi nilai kemudahan perolehan informasi pada aplikasi dalam survei evaluasi luaran penelitian.

Kemudian, terkait dengan kemudahan perolehan informasi dengan antarmuka aplikasi diberikan rata-rata skor sebesar 4,1, dan serupa dengan butir sebelumnya, memiliki distribusi terbanyak pada skala keempat dan kelima (lihat Gambar 4.20). Hal ini menunjukkan bahwa antarmuka aplikasi dapat menyampaikan informasi yang dapat dengan mudah diperoleh oleh pengguna. Cara penyampaian informasi dari perangkat lunak yang menggunakan label sebelum mepresentasikan informasi dan *disclaimer* dapat menjelaskan mengapa secara umum pengguna dapat mendapatkan informasi dengan mudah.

Informasi pada aplikasi ini sangat mudah dicerna
52 responses

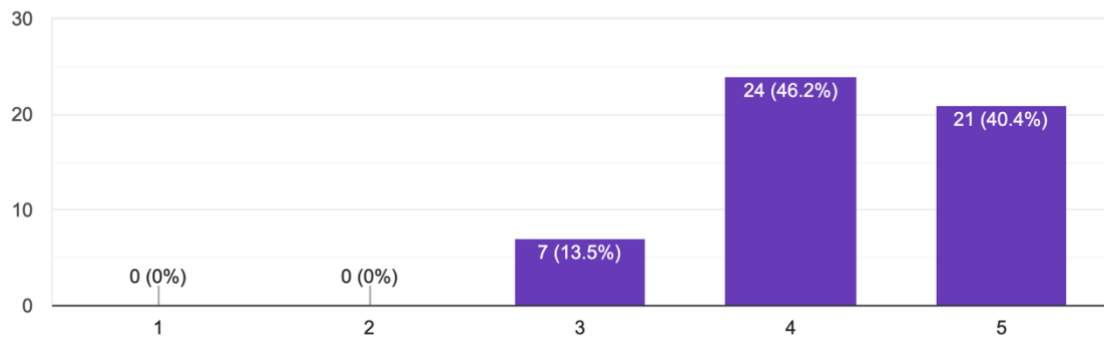


Gambar 4.21 Distribusi nilai kemudahan pencernaan informasi pada aplikasi dalam survei evaluasi luaran penelitian.

Terakhir, pada Gambar 4.21, dapat dilihat untuk kemudahan pencernaan informasi dari aplikasi terdapat hasil yang cukup berbeda, dimana beberapa responden menaruh nilai pada angka 2 dan 3, namun tetap dengan mayoritas skor di angka 4 dan 5. Sehingga, rata-rata untuk aspek ini adalah 4,1, yang mengindikasikan bahwa informasi dari aplikasi dapat dengan mudah dimengerti dan dicerna. Hal ini mungkin dikarenakan perangkat lunak menunjukkan luaran pemodelan topik dalam bentuk visualisasi *nested bubble* dan *line chart* yang intuitif dan mudah dipahami oleh pengguna. Selain itu, salah satu faktor yang dapat mendukung kemudahan pencernaan informasi ini juga dapat dikaitkan dengan label-label yang menjelaskan komponen-komponen dari luaran pemodelan topik.

Dengan aplikasi ini, saya mendapatkan topik-topik yang relevan dengan topik yang saya pilih.

52 responses

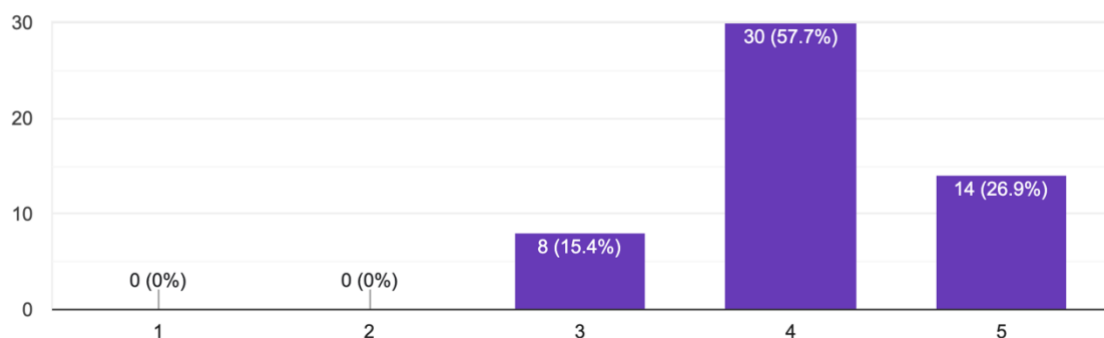


Gambar 4.22 Distribusi nilai relevansi topik pada aplikasi dalam survei evaluasi luaran penelitian.

Dari Gambar 4.22, jika dilihat tentang kerelevanan topik yang dihasilkan oleh perangkat lunak dengan topik yang dipilih oleh responden, sebagian besar responden juga memberikan skor yang tinggi, dengan rata-rata skor 4,2 dari skala 1-5. Ini menunjukkan bahwa topik yang dihasilkan oleh perangkat lunak cukup relevan dengan topik yang dipilih oleh responden. Hal ini dapat dijelaskan melalui pendekatan pemodelan topik baru yang dikemukakan dalam penelitian ini, yang mengekstraksi frasa secara dinamis dari konten artikel menggunakan pola POS dan lalu diproses melalui model LDA, BERT, dan *entity linking*. Melalui pendekatan ini, masing-masing topik yang dihasilkan dilabeli dengan frasa-frasa yang saling terkait dan merujuk ke topik-topik nyata, yang dapat menjelaskan mengapa luaran topik yang dihasilkan relevan dengan kueri pencarian pengguna.

Artikel yang digunakan sesuai dengan label topik yang diberikan dari proses pemodelan topik.

52 responses

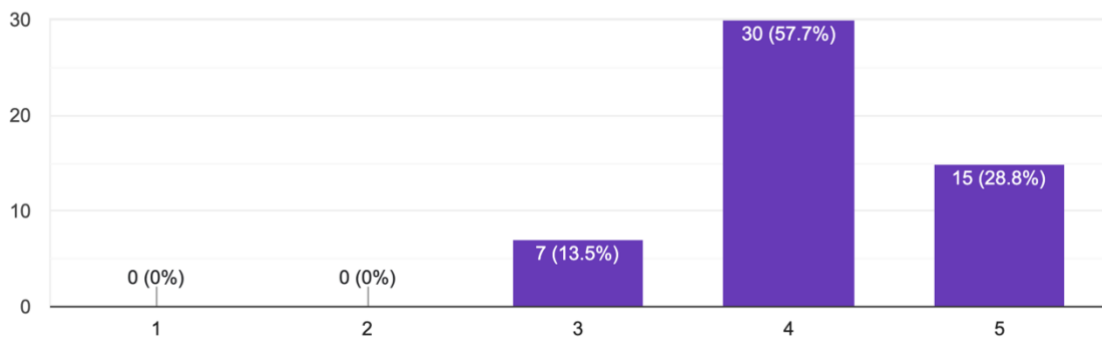


Gambar 4.23 Distribusi nilai keterkaitan artikel dengan label topik pada aplikasi dalam survei evaluasi luaran penelitian.

Sementara itu, untuk evaluasi kualitas kesesuaian pemberian topik pada artikel yang digunakan, rata-rata skor yang diberikan responden adalah 4,3 (lihat Gambar 4.23). Ini menunjukkan bahwa sebagian besar responden menilai bahwa artikel yang digunakan sesuai dengan label yang diberikan dari proses pemodelan topik. Proses pemodelan topik yang menggunakan model LDA pada dasarnya memastikan bahwa setiap topik direpresentasikan oleh kata atau frasa yang sering bermunculan bersama dan mirip melalui proses *Dirichlet*nya. Sehingga, kesesuaian topik pada artikel dapat dikaitkan dengan proses ini, karena konten artikel dapat direpresentasikan oleh topik dengan kata kunci yang serupa.

Hasil luaran dan visualisasi model ini berguna untuk riset saya.

52 responses



Gambar 4.24 Distribusi nilai kegunaan luaran dan visualisasi pada aplikasi dalam survei evaluasi luaran penelitian.

Lalu dari Gambar 4.24, dapat dilihat untuk luaran dan visualisasi yang diberikan oleh perangkat lunak dari hasil pemodelan topik, rata-rata skor yang diberikan adalah 4,4, menunjukkan bahwa visualisasi yang diberikan dapat membantu untuk memberikan gambaran penelitian yang intuitif kepada peneliti. Visualisasi *nested bubble* and *line chart* dapat membantu untuk memberikan gambaran intuitif dari hasil penelitian karena kedua jenis visualisasi ini dapat menyediakan tampilan visual yang sederhana dan mengintegrasikan informasi pemodelan topik dari artikel yang telah dikumpulkan. Secara intuitif, *nested bubble chart* dapat dilihat menyajikan data kuantitatif seperti jumlah dokumen yang diklasifikasikan ke dalam topik tertentu, sementara *line chart* dapat digunakan untuk menunjukkan perubahan dari frekuensi topik dari waktu ke waktu. Dengan menggabungkan kedua jenis visualisasi ini, peneliti dapat dengan cepat dapat mengetahui

topik terkini pada bidang ilmu tertentu dan evolusi tren topik dari waktu ke waktu, yang berguna agar dapat memahami peta penelitian pada topik tertentu.

Berdasarkan hasil survei yang telah dilakukan, terdapat beberapa hal yang dapat disimpulkan. Pertama, sebagian besar responden merasa bahwa antarmuka aplikasi yang telah dikembangkan mudah untuk dinavigasi dan memudahkan mereka untuk mendapatkan informasi dengan mudah. Kedua, informasi yang disajikan dalam aplikasi ini dinilai mudah dicerna oleh sebagian besar responden. Ketiga, sebagian besar responden juga merasa bahwa topik-topik yang didapatkan melalui aplikasi ini relevan dengan topik yang dipilih. Keempat, sebagian besar responden juga merasa bahwa artikel yang digunakan sesuai dengan label topik yang diberikan dari proses pemodelan topik. Terakhir, sebagian besar responden juga merasa bahwa hasil luaran dan visualisasi dari model yang dikembangkan berguna untuk riset mereka.

Bagian berikut akan mendiskusikan sebuah sampel luaran dari perangkat lunak untuk kueri pencarian yang dilakukan oleh salah satu responden, yakni mengenai “*circular economy*”, sebuah topik dalam bidang bisnis dan ekonomika yang menaruh fokus pada penggunaan sumber daya secara berkelanjutan dan efisien yang dapat mengurangi dampak pada lingkungan (Kristoffersen et al., 2021; Yawar & Kuula, 2021). Evaluasi juga akan dilakukan pada luaran pemodelan terhadap kueri pencarian ini dan hasilnya akan dibandingkan dengan luaran topik dari *Open Knowledge Maps*. Proses pencarian dilakukan oleh responden dengan fitur pencarian standar yang mengambil artikel dari semua basis data dengan batasan pengambilan 100 artikel per basis data. Pada Gambar 4.25, dapat dilihat bahwa luaran dari hasil pencarian pada perangkat lunak diperoleh dalam waktu 17.36 menit, dengan total jumlah artikel yang diproses sebanyak 495, 100 dari masing-masing basis data kecuali untuk basis data CORE dimana hanya 95 dokumen yang dapat diperoleh. Adapun jumlah topik yang dihasilkan yakni 18 dengan nilai koherensi topik yang diperoleh sebesar -14.39.

Total Documents Processed: 495

Document Details

Database: arXiv 100 documents

Database: ScienceOpen 100 documents

Database: Emerald 100 documents

Database: Garuda 100 documents

Database: CORE 95 documents

Number of Topics: 18

Coherence Score: -14.39

Process Time: 17.36 minutes

DOWNLOAD PROCESSED DOCUMENTS IN CSV

Gambar 4.25 Tangkapan layar detail parameter pemodelan dari perangkat lunak.

Dalam Tabel 4.6, sampel konten dari lima baris pertama dari dokumen CSV yang dihasilkan oleh perangkat lunak untuk kueri pencarian “*circular economy*” ditampilkan. Seperti yang telah dijelaskan sebelumnya, dokumen CSV menyediakan informasi mengenai artikel yang diproses, seperti judul, abstrak, dan tanggal publikasi penelitian. Selain itu, hasil dari pemodelan topik juga ditampilkan untuk setiap artikel, seperti topik yang ditentukan. Namun, abstrak dan sub frasa yang ditampilkan pada tabel hanya menyajikan sebagian informasi, yang diindikasikan dengan penggunaan elipsis.

Tabel 4.6 Sampel Konten Dokumen CSV untuk Kueri Pencarian *Circular Economy*.

Title	Source	Abstract	Assigned Topic	Published Date	Main Topic	Sub Keyphrases
<i>Circular Economy Case Studies & Snapshots</i>	CORE	<i>The National Zero Waste Council Circular Economy Working Group, in</i>	1	2015-04-04 01:00:00	<i>Circular supply chain</i>	<i>Circular business model, Circular economy business, Circular</i>

Title	Source	Abstract	Assigned Topic	Published Date	Main Topic	Sub Keyphrases
		<i>collaboration with the ...</i>				<i>economy strategy, ...</i>
<i>Circular Economy Snapshot: Frogbox</i>	CORE	<i>FROGBOX is a Vancouver-based moving supply franchise company founded in 2008 ...</i>	8	2014-01-01 00:00:00	<i>Global economy</i>	<i>Economy sector, Market economy, Digital economy, ...</i>
<i>A conceptual framework for circular design</i>	CORE	<i>Design has been recognised in the literature as a catalyst to move away from the traditional model ...</i>	1	2016-09-13 00:00:00	<i>Circular supply chain</i>	<i>Circular business model, Circular economy business, Circular economy strategy, ...</i>
<i>Circular Economy Snapshot: BMW Drivenow</i>	CORE	<i>BMW's views of the future of car ownership influenced its decision to enter the car-sharing business...</i>	3	2015-05-05 01:00:00	<i>Circular migration kota</i>	<i>Migrant garden waste, Livelihood strategy, Circular mobility, Permanent migration thing, ...</i>

Title	Source	Abstract	Assigned Topic	Published Date	Main Topic	Sub Keyphrases
<i>Circular Procurement: The First Step in a Series.</i>	CORE	<i>The Dutch Ministry of Infrastructure & Environment proudly hosted the first International Congress on Circular Procurement ...</i>	1	2016-04-04 01:00:00	<i>Circular supply chain</i>	<i>Circular business model, Circular economy business, Circular economy strategy, ...</i>

Dari sampel konten yang dipaparkan, dapat dilihat bahwa dokumen yang diperoleh oleh perangkat lunak sesuai dengan kueri pencarian responden: “*circular economy*”. Masing-masing artikel dari judulnya merefleksikan bahwa artikel yang diperoleh berbicara tentang *circular economy*, dan jika ditelusuri lebih dalam ke kontennya, dapat dilihat bahwa masing-masing artikel berbicara tentang *zero waste* (pengurangan limbah), penjagaan lingkungan, serta keberlanjutan penggunaan sumberdaya, tiga hal yang sejalan dengan topik *circular economy* secara umum.

Tabel 4.7 menunjukkan hasil pemodelan topik dari perangkat lunak yang digunakan untuk menganalisis teks yang berkaitan dengan *circular economy*. Tabel ini menyajikan topik yang dihasilkan oleh perangkat lunak, label yang merepresentasikan topik tersebut, dan frasa-frasa lain yang dapat merepresentasikan topik terkait.

Tabel 4.7 Luaran Topik Perangkat Lunak untuk Bidang *Circular Economy*

Topik	Label	Frasa
1	<i>Circular supply chain</i>	<i>Circular business model, Circular economy business, Circular economy strategy, Circular economy principle,</i>

Topik	Label	Frasa
		<i>Circular business case, Plastic waste, Waste management, Chain management, Environmental impact economy</i>
2	Kebijakan pemerintah	<i>Food packaging, Pandemic government policy, Student absorption, Community service, Pemerintah desa, Circular letter, Circular letter number, Pembangunan daerah, Research method</i>
3	<i>Circular migration kota</i>	<i>Migrant garden waste, Livelihood strategy, Circular mobility, Permanent migration thing, Adaptation strategy, Permanent mobility, Rural thing, Migrant motivation, Kota household</i>
4	<i>Macro economy</i>	<i>Market economy, Digital economy, Economy existence, Global economy, Abstract economy, National economy production, Economic understanding core, Political economy, Exchange economy</i>
5	<i>Sectoral level trade</i>	<i>Verdoorn law economy, Economy economy, Regional divergence flow, Sectoral level period, Verdoorn law test, Accumulation labour, Portuguese region, Scale prosumer, Flow growth</i>
6	<i>Business cycle</i>	<i>Macro economy existence, Market economy, Digital economy, Global economy, Abstract economy, National economy production, Economy core, Political economy, Economic understanding exchange</i>
7	<i>Terrorism bank</i>	<i>Money laundering, Danger liquidity, Combat finance, Financial institution money, Financial market confidence, Legitimate money loss, Money mix, Circular dpnp year, Country policy</i>
8	<i>Global economy</i>	<i>Economy sector, Market economy, Digital economy, Macro economy core, Abstract economy, National economy production, Economic understanding, Political economy, Exchange economy</i>
9	<i>Local knowledge communication</i>	<i>Local knowledge community, Knowledge research, Economic life society, Information experience, Society</i>

Topik	Label	Frasa
		<i>research, Share information, People information, Policy study, Economic life economy</i>
10	<i>Economy province</i>	<i>Economy sector, Market economy, Digital economy, Economy existence, National economy, Global economy, Abstract economy, Political economy, Sector macro</i>
11	<i>Sectoral level trade</i>	<i>Medical equipment value, Environmental impact paper, Medical waste mouth, Sector maintenance, Waste treatment, Covid crisis, Industrial ecology increase, Medical sector healthcare, Current crisis address</i>
12	<i>Abstract economy</i>	<i>Economy sector, Market economy, National economy, Economy province, Digital economy, Global economy, Economic understanding core, Political economy, Exchange economy</i>
13	<i>Government fuel</i>	<i>Bad fuel economy, Fuel economy, Vehicle traffic, Gap vehicle, Empirical analysis factor, Consumer vehicle, Pair vehicle, Second vehicle, Economy gap</i>
14	Implementasi peraturan	Upaya pemerintah, Penolakan nelayan, Penangkapan ikan, Menteri kelautan, Peraturan menteri, Perikanan nomor, Alat penangkapan, Penggunaan alat, Nelayan pengguna
15	<i>Music industry</i>	<i>Market pop, Market fan, Market economy, Consumer market, Music consumer, Chinese pop music, Large entertainment market, Fan economy, Pop music</i>
16	Pembayaran Indonesia	<i>Otoritas moneter, Bank indonesia, Penawaran valuta, Neraca pembayaran, Uang negara, Kerja indonesia, Mata uang, Fakultas ekonomi, Jakarta bi</i>
17	Keputusan sirkuler	Sirkuler keputusan, Pemegang saham, Kepentingan pemegang, Persetujuan keputusan, Negeri keputusan, Saham perseroan, Saham pt, Pengambilan keputusan, Kredit perbankan

Topik	Label	Frasa
18	<i>Mushroom substrate</i>	<i>Tomato mushroom, Physicochemical analysis tomato, Production tomato, Nutrient source tomato, Bisporus mushroom, Upper spend mushroom, Organic fertilizer, Plant production, Important crop liter</i>

Topik yang dihasilkan meliputi berbagai aspek dari *circular economy*, seperti *circular supply chain* (rantai pasokan yang berkelanjutan), kebijakan pemerintah, *circular migration* kota (migrasi ke *circular economy* di kota), *macro economy* (ekonomi makro), *sectoral level trade* (perdagangan di tingkat sektoral), *business cycle* (siklus bisnis), *terrorism bank* (keamanan bank terhadap terorisme), *global economy* (ekonomi global), *local knowledge communication* (komunikasi pengetahuan lokal), *economy province* (ekonomi provinsi), *sector resource* (sumber daya sektoral), *abstract economy* (ekonomi abstrak), *government fuel* (bahan bakar pemerintah), implementasi peraturan, *music industry* (industri musik), pembayaran di Indonesia, keputusan sirkuler, dan *mushroom substrate* (substrat jamur).

Setiap topik memiliki label yang merepresentasikan topik tersebut, serta frasa-frasa lain yang dapat merepresentasikan topik terkait. Contohnya, untuk topik pertama yaitu *Circular supply chain* memiliki label "*Circular business model*", "*Circular economy business*", "*Circular economy strategy*", "*Circular economy principle*", "*Circular business case*", "*Plastic waste*", "*Waste management*", "*Chain management*", "*Environmental impact economy*" dan frasa lain yang dapat merepresentasikan topik ini antara lain dapat ditemukan dalam frasa-frasa tersebut.

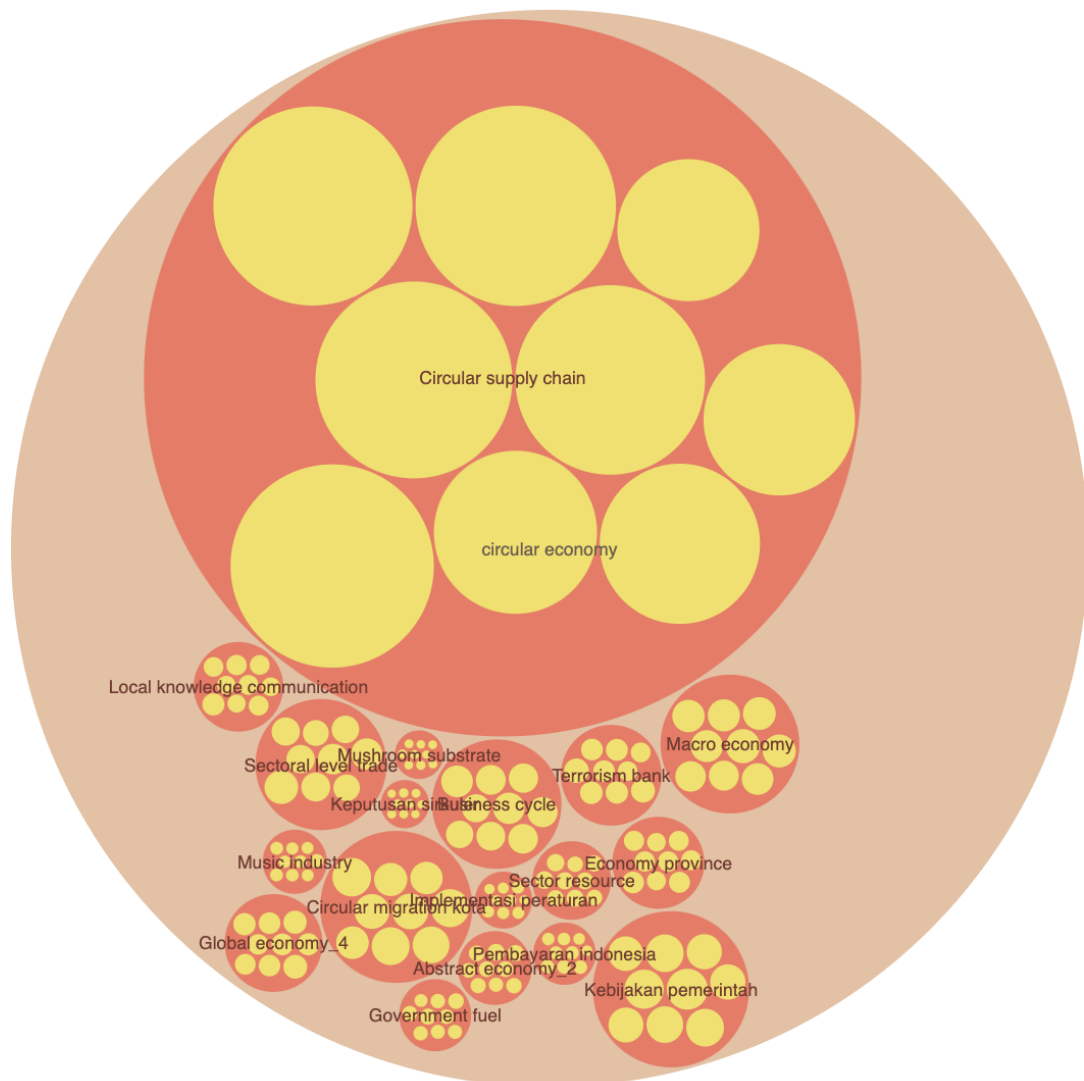
Topik-topik yang dihasilkan dari pemodelan tersebut relevan dengan tema *circular economy*, diantaranya adalah topik tentang *circular supply chain*, kebijakan pemerintah, *circular migration* kota, *macro economy*, serta *business cycle*. Selain itu, topik-topik tersebut juga dikaitkan dengan frasa-frasa yang relevan dengan *circular economy*, seperti *circular business model*, *waste management*, *environmental impact economy*, dan lain sebagainya. Hal ini menunjukkan bahwa proses pemodelan topik yang digunakan dapat memberikan hasil yang akurat dan relevan dengan topik *circular economy*. Selain itu, ini juga menunjukkan kemampuan proses pemodelan topik yang digunakan karena dapat memproses konten multilingual secara bersamaan dan mengelompokkan mereka dalam sebuah topik; contohnya untuk topik kebijakan pemerintah yang direpresentasikan oleh frasa-frasa: "*Food packaging*", "*Pandemic government policy*", "*Student absorption*", "*Community service*",

Pemerintah desa”, “Circular letter”, “Circular letter number”, “Pembangunan daerah”, *Research method*”. Ini menandakan bahwa proses pemodelan topik yang digunakan bahwasanya *language agnostic*, dan berpotensi dapat digunakan tanpa batasan pada bahasa teks yang diproses.

Hasil pemodelan topik yang telah dilakukan pada kueri pencarian ini ditunjukkan pada Gambar 4.26 dan 4.27, yang merupakan visualisasi *nested bubble* dan *line chart*. Visualisasi *nested bubble chart* menunjukkan frekuensi dokumen yang dilabeli pada topik tertentu, dan *line chart* memperlihatkan tren topik dari waktu-ke-waktu.

LDA results for circular economy

Overall Topic Distribution

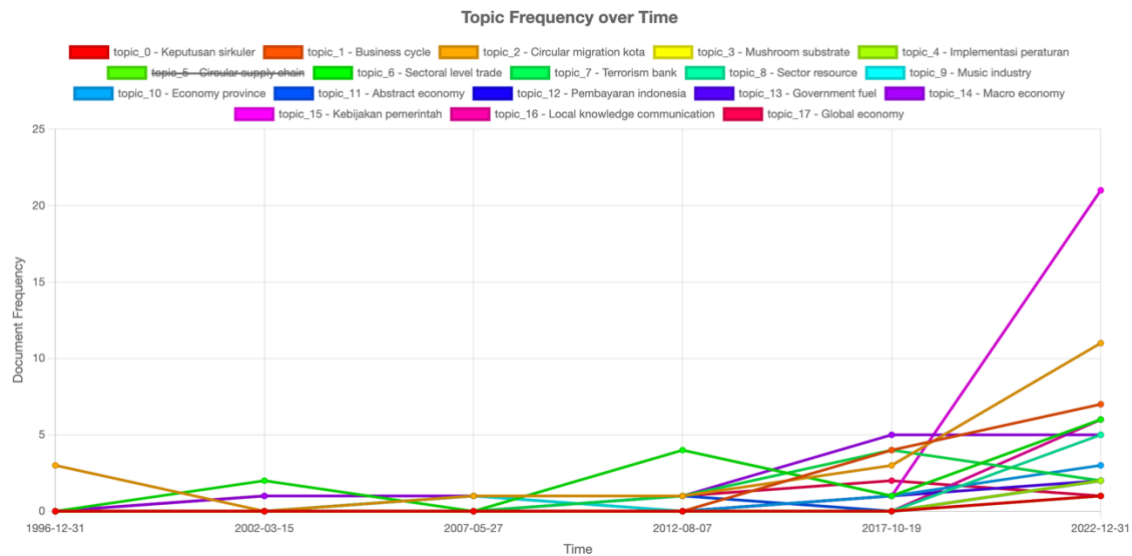


Gambar 4.26 Visualisasi *nested bubble chart* dari hasil pemodelan topik dalam bidang *circular economy*.

Dapat dilihat dari Gambar 4.26, topik *circular supply chain* merupakan topik dengan artikel terbanyak dalam bidang ini. Hal ini secara intuitif masuk akal, karena *circular supply chain* merupakan salah satu konsep utama yang memfasilitasi aliran bahan dan sumber daya dalam lingkaran tertutup melalui tiga langkah utama, yakni: pengurangan, perekayasaan ulang, dan pengolahan ulang sumber daya. Dapat dilihat juga topik-topik yang berbicara tentang kebijakan pemerintah dan migrasi ke ekonomi yang *circular* pada kota juga merupakan topik yang sering dibicarakan, yang dapat menandakan bahwa implementasi *circular economy* membutuhkan dukungan pemerintah melalui realisasi kebijakan-kebijakan yang terkait, serta dapat diimplementasikan di kota-kota.

Namun, dapat dilihat untuk *bubble* topik *mushroom substrate*, keputusan, dan implementasi peraturan sangat kecil. Hal ini menandakan bahwa belum banyak riset yang berbicara tentang hal-hal tersebut. Jika digeneralisasikan, topik *mushroom substrate* berbicara tentang jamur yang merupakan bahan yang sering dibuang, tetapi seharusnya dapat didaur ulang, dan alasan mengapa topik ini mungkin jarang dibicarakan karena sangat teknis dan spesifik ke bahan tertentu. Selain itu, kecilnya *bubble* tentang keputusan dan implementasi peraturan juga dapat dijelaskan dari fokus peneliti untuk mengembangkan peraturan-peraturan untuk mendukung *circular economy*, tetapi peneliti dalam bidang *circular economy* belum terlalu menaruh perhatian pada bagaimana pemerintah dapat melakukan pengambilan keputusan dan mengimplementasikan peraturan ini. Hal Ini juga dapat diartikan sebagai sebuah *gap* dalam literatur, dimana terdapat kurangnya penelitian dalam topik tersebut.

Lalu, *line chart* yang ditunjukkan pada Gambar 4.27, menunjukkan tren frekuensi dokumen untuk setiap topik dari waktu-ke-waktu.



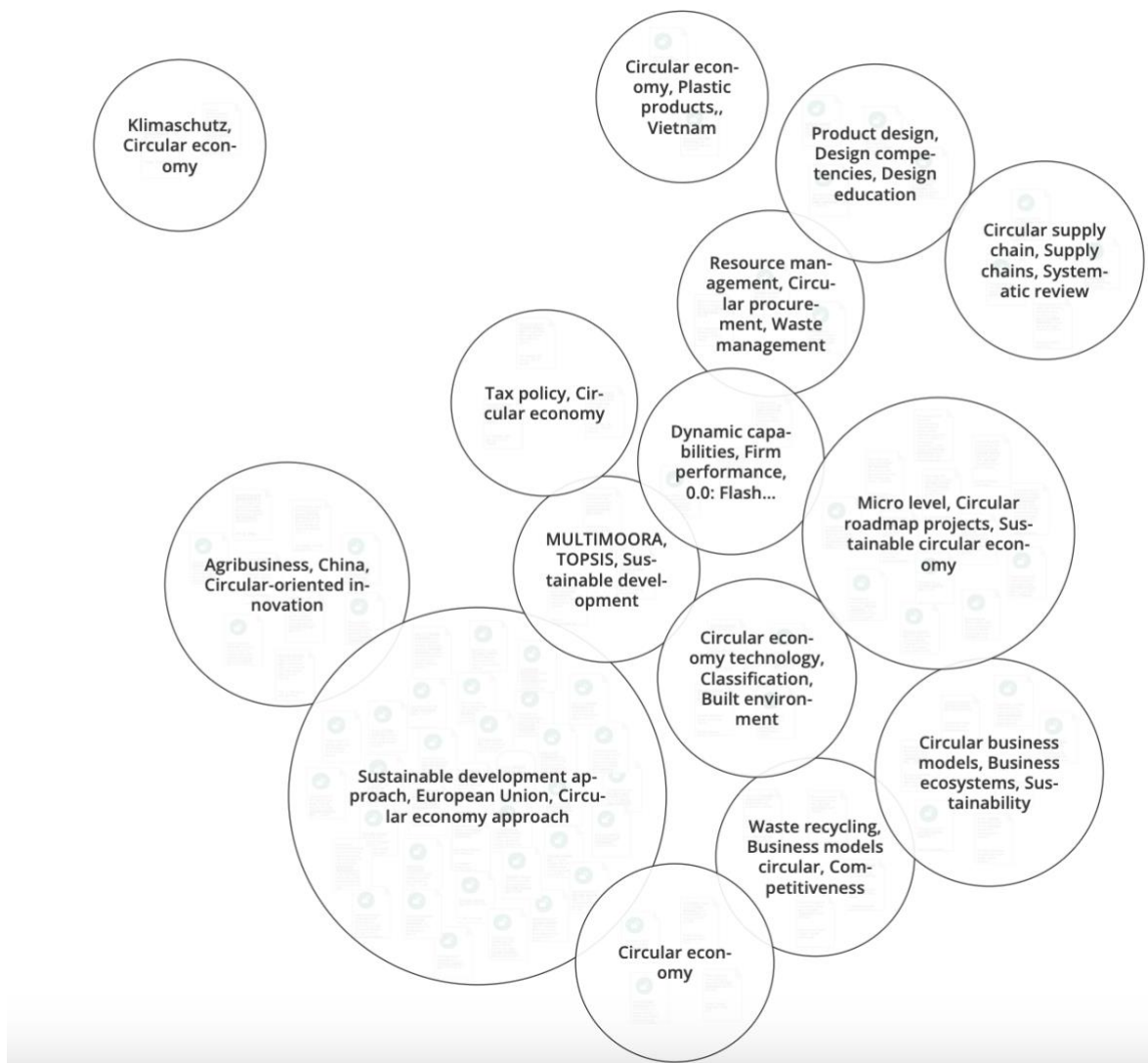
Gambar 4.27 Visualisasi *line chart* dari hasil pemodelan topik dalam bidang *circular economy*.

Dari gambar di atas, dapat diinferensikan bahwa terdapat kenaikan yang signifikan dalam penelitian mengenai kebijakan pemerintah untuk *circular economy*. Hal ini menunjukkan bahwa minat dalam topik ini semakin meningkat dan mendapat perhatian yang lebih dari para peneliti dan pemerintah. Dari visualisasi tersebut juga, dapat dilihat bahwa sebelum tahun 2017, mayoritas penelitian masih berfokus pada tema-tema umum, seperti *sectoral level trade*, *macro*, dan *global economy*. Hal ini menarik untuk diteliti lebih lanjut untuk mengetahui alasan mengapa ada ledakan dalam penelitian mengenai kebijakan pemerintah, serta topik-topik lainnya seperti migrasi ke ekonomi *circular* pada kota pada tahun 2017 dan seterusnya.

Untuk membandingkan sampel luaran yang telah dipaparkan dengan hasil *Open Knowledge Maps*, sebuah peta penelitian dibuat dengan *Open Knowledge Maps* dengan kata kunci yang sama, yakni “*circular economy*”, dengan pilihan basis data yang digunakan adalah BASE karena merupakan opsi dari perangkat lunak *Open Knowledge Maps* yang menginkorporasi semua bidang ilmu. Hasil luaran visualisasi dari *Open Knowledge Maps* dapat dilihat pada Gambar 4.28 yang menunjukkan peta penelitian yang dihasilkan.

Knowledge Map of **circular economy**

100 most relevant documents Data source: BASE Until 25 Feb 2023 Document types Data quality More information



Gambar 4.28 Visualisasi peta penelitian dari *Open Knowledge Maps* dalam bidang *circular economy* (Open Knowledge Maps, 2023).

Dari peta penelitian yang dihasilkan oleh *Open Knowledge Maps* pada bidang *circular economy*, dapat dilihat bahwa topik *sustainable development approach*, *European union*, dan *circular economy approach* merupakan tema terbesar dalam bidang ini. Selain itu, topik-topik terkait model bisnis sirkuler, serta pengaturan sumber daya dan sampah juga menjadi topik utama dalam bidang *circular economy*. Hal ini cukup berbeda dengan hasil dari luaran penelitian, dimana topik-topik utama membicarakan tentang *circular supply chain* dan kebijakan pemerintah dalam mewujudkan ekonomi yang sirkuler.

Walaupun begitu, terdapat juga beberapa kemiripan dari topik yang dihasilkan oleh *Open Knowledge Maps* dengan luaran penelitian. Semisal, topik *circular supply chain*, pengaturan sampah dan plastik (*waste management, plastic products, waste recycling*), serta peraturan-peraturan pemerintah (*tax policy* dan kebijakan pemerintah) muncul pada luaran kedua perangkat lunak. Dari sisi luaran penelitian, hal ini dapat mengindikasikan bahwa hasil penelitian cukup akurat dan relevan, karena *Open Knowledge Maps* telah banyak digunakan dalam penelitian dan dianggap sebagai sumber tepercaya untuk mengidentifikasi topik penelitian mutakhir.

Adapun juga topik-topik yang berbeda dari masing-masing perangkat lunak. Dari hasil *Open Knowledge Maps*, terdapat topik *sustainable development, agribusiness, circular-oriented-innovation*, dan *product design* yang mana topik-topik tersebut tidak dihasilkan oleh luaran penelitian. Namun, juga terdapat topik yang ada dari luaran penelitian tapi tidak ada dari hasil *Open Knowledge Maps*, seperti migrasi dan komunikasi ekonomi sirkuler ke kota dan komunitas, *mushroom substrate*, serta keputusan dan implementasi peraturan dan kebijakan pemerintah. Selain itu, luaran penelitian ini juga menghasilkan topik-topik *circular economy* berbahasa Indonesia selain topik berbahasa Inggris. Perbedaan hasil topik dari kedua perangkat lunak mungkin disebabkan oleh penggunaan basis data akademis yang berbeda, dikarenakan *Open Knowledge Maps* hanya mendukung PubMed dan BASE; sedangkan basis data yang digunakan pada luaran penelitian mendukung lima basis data akademis, yakni CORE, arXiv, ScienceOpen, Emerald, dan Garuda.

Pada *Open Knowledge Maps*, terlihat pada beberapa lingkaran, terdapat topik “*circular economy*” yang mengulang, padahal, topik tersebut adalah kata kunci pencarian atau bidang ilmu yang ingin diteliti, sehingga hasil topik ini kontra intuitif dari peta penelitian yang dihasilkan. Pada luaran penelitian ini, topik “*circular economy*” tidak muncul berulang kali. Hal ini dapat dijelaskan dari metode filtrasi frekuensi frasa yang terlalu banyak yang diaplikasikan pada penelitian.

Selain itu, luaran penelitian menyediakan visualisasi *nested bubble* dan *line chart* yang memungkinkan pengguna untuk memvisualisasikan tidak hanya topik terkini, tetapi juga kesenjangan dan tren dalam evolusi penelitian. Fitur-fitur ini dapat memberikan pengguna pemahaman yang lebih komprehensif tentang kecanggihan di bidang tertentu, dan dapat membantu mengidentifikasi area penelitian baru atau potensi kolaborasi. Dari hasil perbandingan diatas, dapat disimpulkan bahwa luaran penelitian dapat menghasilkan topik yang akademik, dibuktikan dari kemiripan antara hasil luaran penelitian dan *Open Knowledge Maps* yang serupa. Selain itu, luaran penelitian juga menawarkan beberapa fitur

tambahan yang dapat meningkatkan pemahaman pengguna tentang topik terkini, evolusi riset, serta memberikan representasi penelitian yang akurat dan komprehensif.

Dari sampel kueri pencarian pengguna yang telah dipaparkan, serta perbandingan yang telah dilakukan dengan perangkat lunak yang telah banyak digunakan oleh peneliti lainnya, dapat dilihat bahwa luaran penelitian dapat menghasilkan topik *state-of-the-art* dalam sebuah topik penelitian, dan pengaplikasiannya pun dapat diuji coba pada topik-topik riset pada bidang ilmu serta bahasa apapun dikarenakan sifat pemrosesan yang *language agnostic* dan *context independent*. Hal ini menunjukkan bahwa proses pemodelan topik yang dipaparkan pada penelitian ini dapat digunakan pada berbagai bidang dan dapat memberikan hasil yang akurat serta relevan dalam menganalisis tren dalam penelitian.

Mempertimbangkan hasil survei serta hasil evaluasi lanjutan pada sampel kueri pencarian pengguna pada perangkat lunak, dapat disimpulkan bahwa luaran penelitian yang telah dikembangkan dapat membantu peneliti dalam menemukan *state-of-the-art* dari sebuah topik penelitian dan memberikan wawasan tentang pengembangan terkini secara otomatis. Perangkat lunak yang dihasilkan dapat digunakan oleh peneliti secara umum di berbagai bidang penelitian untuk memperoleh topik *state-of-the-art* dalam riset dan dapat juga digunakan untuk menemukan kesenjangan dalam riset.

BAB 5

Kesimpulan dan Saran

5.1 Kesimpulan

Upaya untuk mengembangkan ilmu pengetahuan merupakan upaya yang semestinya semakin mudah untuk dilakukan, namun malah menjadi semakin menantang dengan adanya terlalu banyak informasi secara digital, tanpa bantuan alat yang dapat memproses informasi-informasi tersebut untuk memberikan wawasan pada peneliti. Terdapat penelitian sebelumnya yang telah menawarkan berbagai cara untuk membantu peneliti dalam mengolah informasi ini untuk merumuskan peta penelitian, namun pendekatan yang ada masih belum dapat melakukannya secara otomatis, dan luaran hasil yang diperoleh juga belum intuitif karena hanya dibatasi pada data bibliometris yang minim.

Berdasarkan hasil penelitian yang telah dilakukan dan masalah yang telah dipaparkan, dapat disimpulkan bahwa perangkat lunak yang dikembangkan dapat membantu peneliti dalam menemukan *state-of-the-art* dari sebuah topik penelitian. Hasil penelitian ini menjawab rumusan masalah yang diajukan dan memberikan solusi yang dapat membantu peneliti menemukan *state-of-the-art* dari sebuah topik penelitian dan memberikan wawasan tentang pengembangan terkini secara otomatis. Temuan-temuan pokok dari penelitian ini adalah bahwa perangkat lunak yang dikembangkan mampu mengumpulkan data dari berbagai basis data akademis secara otomatis, mengaplikasikan model pembelajaran mesin dan analisis teks untuk menemukan topik pada penelitian, dan menyajikan luaran topik dalam bentuk visualisasi yang memudahkan interpretasi.

Penggunaan model pembelajaran mesin LDA, serta ekstraksi frasa secara otomatis menggunakan pola POS dan model BERT terbukti dapat membantu dalam menemukan topik yang terkait dengan *state-of-the-art* penelitian, sedangkan metode penyaringan topik berbasis *entity linking* dapat memastikan bahwa luaran topik yang sesuai, akurat, dan relevan dengan topik penelitian. Sifat model LDA yang agnostik terhadap bahasa pada teks yang diberikan, disertai dengan model BERT yang multilingual juga menunjukkan kemampuan luaran penelitian untuk mensintesis dan menemukan topik pada konten artikel lintas bahasa. Visualisasi topik yang disajikan dalam bentuk *nested bubble chart* dan *line chart* juga memudahkan peneliti dalam memahami tren dan kesenjangan dalam penelitian. Implikasi dari hasil penelitian ini adalah bahwa proses pemodelan topik dan perangkat lunak yang dikembangkan dapat membantu peneliti dalam menemukan *state-of-the-art* penelitian dan

memberikan wawasan tentang pengembangan terkini secara otomatis. Luaran penelitian ini juga dapat digunakan oleh peneliti di berbagai bidang penelitian untuk memperoleh informasi terkini dan menemukan kesenjangan dalam penelitian.

5.2 Saran

Penelitian ini memiliki beberapa keterbatasan yang dapat dikembangkan pada penelitian selanjutnya. Pertama, perangkat lunak hanya dapat mengumpulkan data dari beberapa basis data akademis yang terpilih, yakni CORE, arXiv, ScienceOpen, Emerald, dan Garuda, sehingga tidak semua data penelitian dapat diperoleh. Kedua, terdapat kemungkinan luaran topik masih mengandung topik yang tidak relevan karena tidak semua data dapat difilter dengan baik menggunakan basis data pengetahuan yang tersedia. Selain itu, perangkat lunak saat ini memakan waktu yang cukup lama (sekitar 4 menit) untuk menganalisis teks dan menghasilkan luaran.

Arah penelitian selanjutnya yang dapat dilakukan adalah mengembangkan perangkat lunak untuk mengumpulkan data dari sumber lain selain basis data akademis, semisal dari Google Scholar, ScienceDirect, dan basis data akademis lainnya. Selain itu, penelitian selanjutnya dapat mengeksplorasi pendekatan lainnya untuk menghasilkan topik yang lebih akurat dan relevan secara otomatis, semisal dengan menggunakan konten *full-text* dari artikel yang diperoleh disertai dengan *large language models* (LLM) yang dilatih pada teks akademis. Penelitian selanjutnya juga dapat mengevaluasi efektivitas perangkat lunak ini dalam meningkatkan efisiensi pencarian state-of-the-art bagi peneliti, menggunakan uji validitas dan reliabilitas yang lebih ketat, serta mengevaluasi kemampuan perangkat lunak dalam menyajikan hasil penelitian secara lebih terperinci dan intuitif.

Daftar Pustaka

- Abbas, A., & Mahsen Abdulrahman, M. (2023). About Scholarly Works in the Field of Islamic Economic Law: A Visualization of Related Topics. *Jurnal Hukum Ekonomi Syariah*, 2(1), 46–55. <https://doi.org/10.46870/milkiyah.v2i1.267>
- Ali, C. B., Wang, R., & Haddad, H. (2015). A two-level keyphrase extraction approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9042, 390–401. https://doi.org/10.1007/978-3-319-18117-2_29
- Al-Ramahi, M., El-Gayar, O., & Liu, J. (2016). Discovering design principles for persuasive systems: A grounded theory and text mining approach. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2016-March*, 3074–3083. <https://doi.org/10.1109/HICSS.2016.387>
- Ameen, K., Batool, S. H., & Naveed, M. A. (2019). Difficulties novice LIS researchers face while formulating a research topic. *Information Development*, 35(4), 592–600. <https://doi.org/10.1177/0266666918774875>
- Angkawijaya, Y. (2022). PEMETAAN PENELITIAN TERHADAP TOPIK ADVERGAMING. *AKSA JURNAL DESAIN KOMUNIKASI VISUAL*, 6(1), 846–858. <http://aksa.stsrdvisi.ac.id>
- Arminditya Fajri Akbar, Harry Budi Santoso, Putra, P. O. H., & Yudhoatmojo, S. B. (2021). User Perception Analysis of Online Learning Platform “Zenius” During the Coronavirus Pandemic Using Text Mining Techniques. *Jurnal Sistem Informasi*, 17(2), 33–47. <https://doi.org/10.21609/jsi.v17i2.1065>
- Arun, R., Suresh, V., Veni Madhavan, C. E., & M. N., N. M. (2010). *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*. https://doi.org/10.1007/978-3-642-13657-3_43
- arXiv. (2021, May 24). *arXiv API Access*. ArXiv API Access. <https://arxiv.org/help/api>
- Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, 16(1), 101224. <https://doi.org/10.1016/j.joi.2021.101224>

- Bellaouar, S., Bellaouar, M. M., & Ghada, I. E. (2021). Topic modeling: Comparison of LSA and LDA on scientific publications. *ACM International Conference Proceeding Series*, 59–64. <https://doi.org/10.1145/3456146.3456156>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.
- Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).
- Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, 6(3). <https://doi.org/10.1371/journal.pone.0018029>
- Breitinger, C., Wortner, P., Gipp, B., & Reiterer, H. (2019). “Too late to collaborate”: Challenges to the discovery of in-progress research. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2019-June*, 134–137. <https://doi.org/10.1109/JCDL.2019.00028>
- Carlisle, J. P., Cunningham, S. W., Nayak, A., & Porter, A. (1999). *RELATED PROBLEMS OF KNOWLEDGE DISCOVERY*.
- Chen, D. T. V., Wang, Y. M., & Lee, W. C. (2016). Challenges confronting beginning researchers in conducting literature reviews. *Studies in Continuing Education*, 38(1), 47–60. <https://doi.org/10.1080/0158037X.2015.1030335>
- CORE. (n.d.-a). *About CORE*. <https://core.ac.uk/about#contact>
- CORE. (n.d.-b). *CORE API (3.0.0)*. <https://api.core.ac.uk/docs/v3>
- Daenekindt, S., & Huisman, J. (2020). Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018. In *Higher Education* (Vol. 80, Issue 3, pp. 571–587). Springer. <https://doi.org/10.1007/s10734-020-00500-x>
- Davis, J., Mengersen, K., Bennett, S., & Mazerolle, L. (2014). Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus*, 3(1). <https://doi.org/10.1186/2193-1801-3-511>
- Delvechia Agustin, L., Noor Istiqomah, I., Kampus Lumajang, K., & Jember, U. (2022). Manajemen Kadar Glukosa Darah Pasien Diabetes Mellitus Tipe 2 Melalui Puasa Ramadhan. *Jurnal Kesehatan Komunitas Indonesia*, 2(1), 78–91. <https://ebsina.or.id/journals/index.php/jkki>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <http://arxiv.org/abs/1810.04805>
- Dodig Crnkovic, G. (2002, April). *Scientific Methods in Computer Science*.
- Farra Fazria, N., & Dhamayanti, I. (2021). A Literature review on the Identification of Variables for Measuring Hospital Efficiency in the Data Envelopment Analysis (DEA). *Unnes Journal of Public Health*, 10(1). <https://doi.org/10.15294/ujph.v0i0.34141>
- Felizardo, K. R., Macdonell, S. G., Mendes, E., & Maldonado, J. C. (2012). A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews. In *Journal of Software* (Vol. 7, Issue 2, pp. 450–461). <https://doi.org/10.4304/jsw.7.2.450-461>
- Francesca Chandra, A., & Nadjib, M. (2023). Digital Marketing in The Hospital: A Scoping Review. *Journal of World Science*, 2(1), 1325–1332. <https://doi.org/10.58344/jws.v2i1.178>
- Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122(1359), 108–111. <https://doi.org/10.1126/science.122.3159.108>
- Garuda - Garba Rujukan Digital. (2018). <https://garuda.kemdikbud.go.id/>
- Geetha, M. P., & Karthika Renuka, D. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks*, 2(July), 64–69. <https://doi.org/10.1016/j.ijin.2021.06.005>
- Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. In *Scientometrics* (Vol. 111, Issue 2, pp. 981–998). Springer Netherlands. <https://doi.org/10.1007/s11192-017-2296-z>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, 10, 5228–5235. <https://doi.org/10.1073/pnas.0307752101>
- Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194, 130–150. <https://doi.org/10.1016/j.artint.2012.04.005>
- Han, S., Gao, N., Guo, X., & Shan, Y. (2021). Incorporating Common Knowledge and Specific Entity Linking Knowledge for Machine Reading Comprehension. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence*

- and *Lecture Notes in Bioinformatics*), 12817 *LNAI*, 546–558.
https://doi.org/10.1007/978-3-030-82153-1_45
- Handayani, N., Taufiq, R., & Mayatopani, H. (2022). Development of Artificial Intelligence in Internet of Things: Review. *CYBERNETICS*, 6(01), 34–42.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. In *Nature* (Vol. 585, Issue 7825, pp. 357–362). Nature Research. <https://doi.org/10.1038/s41586-020-2649-2>
- Haseeb Zia, M., Asif, A., & Ur Rehman, A. (2021). Nutritional Intervention Based on Moringa Leaf Flour to Prevent Stunting in First 1000 Days of Life (1000 HPK): Review Articles. In *Systematic Reviews in Pharmacy* (Vol. 12, Issue 1).
- Hassan, S., Al-Augby, M., Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA Topic Modeling Classification: Comparison study on E-books Sentiment Analysis on Social Media View project LSA & LDA Topic Modeling Classification: Comparison study on E-books. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1). <https://doi.org/10.11591/ijeecs.v19.i1.pp%25p>
- Holzinger, A. (2019). Big data calls for machine learning. *Encyclopedia of Biomedical Engineering*, 1–3, 258–264. <https://doi.org/10.1016/B978-0-12-801238-3.10877-3>
- Jahan, N., Naveed, S., Zeshan, M., & Tahir, M. A. (2016). How to Conduct a Systematic Review: A Narrative Literature Review. *Cureus*. <https://doi.org/10.7759/cureus.864>
- Jasimuddin, S. M., Connell, C., & Klein, J. H. (2005). The challenges of navigating a topic to a prospective researcher: the case of knowledge management research. *Management Research News*, 28(1), 62–76. <https://doi.org/10.1108/01409170510784715>
- Kamogelo, M., & Lucky, M. T. (2022). Knowledge production and information dissemination: challenges faced by new and upcoming researchers. *EUREKA: Social and Humanities*, 3, 60–69. <https://doi.org/10.21303/2504-5571.2022.002446>
- Karampela, S., Andreopoulos, A., & Koutsouris, A. (2021). “Agro”, “Agri”, or “Rural”: The Different Viewpoints of Tourism Research Combined with Sustainability and Sustainable Development. *Sustainability (Switzerland)*, 13(17). <https://doi.org/10.3390/su13179550>
- Open Knowledge Maps. (2023). *Knowledge Map for research on circular economy*. Open Knowledge Maps. <https://openknowledgemaps.org/map/705b238cac20818cc47872d1dd740e7b>

- Koloski, B., Pollak, S., Škrlić, B., & Martinc, M. (2022). *Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?* <http://arxiv.org/abs/2202.06650>
- Kraker, P., Kittel, C., & Enkhbayar, A. (2016). Open Knowledge Maps: Creating a Visual Interface to the World's Scientific Knowledge Based on Natural Language Processing. *027.7 Zeitschrift Für Bibliothekskultur*, 4(2), 98–103. <https://doi.org/10.12685/027.7-4-2-157>
- Kraker, P., Schramm, M., & Kittel, C. (2019). Open knowledge maps: Visual discovery based on the principles of open science. *VOEB-Mitteilungen*, 72(2), 460–477. <https://doi.org/10.31263/voebm.v72i2.3202>
- Kristoffersen, E., Mikalef, P., Blomsma, F., & Li, J. (2021). Towards a business analytics capability for the circular economy. *Technological Forecasting and Social Change*, 171(June), 120957. <https://doi.org/10.1016/j.techfore.2021.120957>
- Landauer, T. K., & Dutnais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. In *Psychological Review* (Vol. 1, Issue 2).
- Lee, C.-H., Liu, C.-L., Trappey, A. J. C., Mo, J. P. T., & Desouza, K. C. (2021). Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. *Advanced Engineering Informatics*, 50, 101428. <https://doi.org/10.1016/j.aei.2021.101428>
- Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science Journal*, 9.
- Li, A. A. S., Trappey, A. J. C., Trappey, C. v., & Fan, C. Y. (2019, October 9). *E-discover State-of-the-art Research Trends of Deep Learning for Computer Vision*.
- Liu, H., Jiang, Y., Fan, H., Wang, X., & Zhao, K. (2021). Visualization Analysis of Knowledge Network Research Based on Mapping Knowledge. *Journal of Signal Processing Systems*, 93(2–3), 333–344. <https://doi.org/10.1007/s11265-020-01595-2>
- Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. In *ACL and AFNLP*.
- Mandaka, M., Nuryanti, W., & Titisari, D. (2022). Exploring the Potential Tourism Objects in the Small City of Heritage Lasem. *Journal of Humanities and Social Sciences Studies*, 4(4), 36–47. <https://doi.org/10.32996/jhsss>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd*

- Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.
- Mathur, V., Stavrakas, Y., & Singh, S. (2016). Intelligence analysis of Tay Twitter bot. *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, 231–236. <https://doi.org/10.1109/IC3I.2016.7917966>
- Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human Computer Studies*, 67(9), 716–754. <https://doi.org/10.1016/j.ijhcs.2009.05.004>
- Meschenmoser, P., Meuschke, N., Hotz, M., & Gipp, B. (2016). Scraping scientific web repositories: Challenges and solutions for automated content extraction. *D-Lib Magazine*, 22(9–10). <https://doi.org/10.1045/september2016-meschenmoser>
- Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.
- Montani, I., Honnibal, M., Honnibal, M., Landeghem, S. van, Boyd, A., Peters, H., McCann, P. O., geovedi, jim, O'Regan, J., Samsonov, M., Altinok, D., Orosz, G., Kok, D. de, Kristiansen, S. L., Bournhonesque, R., Kannan, M., Miranda, L., Baumgartner, P., Edward, ... Bozek, S. (2022). *spaCy: Industrial-strength Natural Language Processing in Python*. <https://doi.org/10.5281/ZENODO.7310816>
- Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2015). Topic modelling for qualitative studies. *Journal of Information Science*, 43(1). <https://doi.org/10.1177/016555151515617393>
- Oliveira, O., Silva, F., Juliani, F., Barbosa, L., & Nunhes, T. (2019). *Bibliometric Method for Mapping the State-of-the-Art and Identifying Research Gaps and Trends in Literature: An Essential Instrument to Support the Development of Scientific Projects*. IntechOpen. <https://doi.org/10.5772/intechopen.85856>
- Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). <http://scikit-learn.sourceforge.net>.
- Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.
- Rabby, G., Azad, S., Mufti, M., Zamli, K. Z., & Mostafizur Rahman, M. (2018). A Flexible Keyphrase Extraction Technique for Academic Literature. *Procedia Computer Science*, 135, 553–563. <https://doi.org/10.1016/j.procs.2018.08.208>

- Rantung, D. I., Sysca Oroh, N., & Prayanthi, I. (2022). JRAM (Jurnal Riset Akuntansi Multiparadigma) Literature Review: Multiparadigm Accounting Research. *Jurnal Riset Akuntansi Multiparadigma*, 9(2), 28–41. <https://jurnal.uisu.ac.id/index.php/JRAM>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, 3982–3992. <https://doi.org/10.48550/arxiv.1908.10084>
- Resti, J., & Selva Jumeilah, F. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal Rekayasa Sistem Dan Informasi Teknologi*, 1(1), 19–25. <http://jurnal.iaii.or.id>
- Revere, D., Fuller, S. S., Bugni, P. F., & Martin, G. M. (2003). A new system to support knowledge discovery: Telemakus. *Proceedings of the ASIST Annual Meeting*, 40, 52–58. <https://doi.org/10.1002/meet.1450400106>
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. <https://doi.org/10.1145/2684822.2685324>
- Roshdi, A., & Roohparvar, A. (2015). Review: Information Retrieval Techniques and Applications. In *International Journal of Computer Networks and Communications Security* (Vol. 3, Issue 9). www.ijcnscs.org
- Rumbut, J., Fang, H., & Wang, H. (2020). Topic modeling for systematic review of visual analytics in incomplete longitudinal behavioral trial data. *Smart Health*, 18, 100142. <https://doi.org/10.1016/j.smhl.2020.100142>
- S., V., J. Ilamathi, & Nithya. (2015). Preprocessing Techniques for Text Mining - An Overview. In *International Journal of Computer Science & Communication Networks* (Vol. 5, pp. 7–16).
- Santos, C. T. dos. (2019). *How to write a good state of the art: should it be the first step of your thesis?*
- Schwab, L. (2021, July 2). *arxiv.py*. <https://github.com/lukasschwab/arxiv.py/blob/master/README.md>
- ScienceOpen. (2017, December 23). *What is ScienceOpen?* <https://about.scienceopen.com/what-is-scienceopen/#more-38>
- Seng Hansen. (2022). IDENTIFIKASI JENIS BAHAYA DAN PARAMETER PENILAIAN BAHAYA PADA PEKERJAAN KONSTRUKSI. *PADURAKSA: Jurnal Teknik Sipil Universitas Warmadewa*, 11(1), 94–102. <https://doi.org/10.22225/pd.11.1.4356.94-102>

- Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 443–460. <https://doi.org/10.1109/TKDE.2014.2327028>
- Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2012). Syntactic Dependency-based N-grams as Classification Features. *Proceedings of the 11th Mexican International Conference on Advances in Computational Intelligence*. https://doi.org/10.1007/978-3-642-37798-3_1
- Smatana, M., & Butka, P. (2019). TopicAE: A Topic Modeling Autoencoder. In *Acta Polytechnica Hungarica* (Vol. 16, Issue 4).
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Southern Denmark University. (2021). *State of the art*. <https://www.sdu.dk/en/bibliotek/studerende/opgaveskrivning/state+of+the+art>
- Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451. <http://arxiv.org/abs/1612.03975>
- Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), 2464–2476. <https://doi.org/10.1002/asi.23596>
- Thornton, S. (2022). *Karl Popper*. The Stanford Encyclopedia of Philosophy (Winter 2022 Edition). <https://plato.stanford.edu/archives/win2022/entries/popper/>
- University of Cambridge. (n.d.). *Emerald Insight*. <https://www.alumni.cam.ac.uk/benefits/journals-and-online-resources/emerald-insight>
- University of the West of Scotland. (2022, September 12). *A-Z Databases: Emerald Insight: Library Guide*. <https://uws-uk.libguides.com/emerald>
- Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic Modeling: Latent Semantic Analysis for the Social Sciences. *Social Science Quarterly*, 99(5), 1665–1679. <https://doi.org/10.1111/ssqu.12528>
- Wahono, R. S. (2015). A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks. *Journal of Software Engineering*, 1(1).

<https://journal.ilmukomputer.org/index.php?journal=jse&page=article&op=view&path%5B%5D=47&path%5B%5D=23>

- Wai, F. I. (2022). Deskripsi Review Kinerja Karyawan Pada Perusahaan Di Masa Pandemi Covid 19. *Jurnal Ekonomi Dan Bisnis Digital (EKOBIL)*, 1(2), 95–100. <https://jurnal.smartindo.org/index.php/ekobil>
- Wan, X., & Xiao, J. (2008). *CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction*. Manchester.
- Yawar, S. A., & Kuula, M. (2021). Circular economy and second-hand firms: Integrating ownership structures. *Cleaner Logistics and Supply Chain*, 2(October), 100015. <https://doi.org/10.1016/j.clscn.2021.100015>
- Zervanou, K. (2010). *UvT: The UvT Term Extraction System in the Keyphrase Extraction task*. Association for Computational Linguistics. <http://www.ai.mit.edu/>
- Zhai, K., & Boyd-Graber, J. (2013). *Online Latent Dirichlet Allocation with Infinite Vocabulary*.

LAMPIRAN A

A. Hasil Survei Luaran Penelitian

Timestamp	Email Address	Nama (Inisial)	Fakultas	Minat Riset	A1	A2	A3	PT1	PT2	PT3
11/29/2022 20:14:35	l*****a@ukrida.a c.id	ls	Psikologi dan Ilmu Sosial Budaya	Dinamika kepribadian	5	5	5	5	5	5
12/5/2022 11:08:46	f***n@uii.ac.id	AFH	Teknologi Industri	Natural language processing	4	4	5	4	4	5
12/5/2022 19:57:10	h***n@uii.ac.id	hp	Teknologi Industri	Software engineering	4	3	2	4	4	3
12/7/2022 8:21:24	m***** **h@gmail.com	MFNM	Teknologi Industri	Humanitarian logistics, sustainable supply chain	3	4	3	4	3	3
12/7/2022 9:13:49	e***a@uii.ac.id	E	Teknologi Industri	Medis	3	4	4	4	4	4
12/7/2022 9:57:59	F *****7@g mail.com	FRM	Teknologi Industri	Sains dan teknologi	5	5	5	5	5	5
12/7/2022 10:40:56	s*****i@uii. ac.id	I***a	Teknologi Industri	Circular Economy	5	4	4	3	3	3
12/8/2022 6:33:21	k*****o@ uii.ac.id	kh	Teknologi Industri	Sistem Informasi Enterprise	4	4	3	4	3	4

1/4/2023 18:08:11	r*****1@gmail.com	RD	Bisnis dan Ekonomika	e-commerce	4	4	5	3	4	4
1/3/2023 19:09:08	t*****o@gmail.com	Teo	Bahasa dan Seni	-	3	3	4	4	4	3
1/5/2023 21:39:39	i*****i@gmail.com	I*****i	Teknik Sipil dan Perencanaan	manajemn proyek	4	4	5	4	5	4
1/3/2023 16:43:35	s*****a@gmail.com	I*****n	Pendidikan	PG SD	4	4	4	4	5	4
1/5/2023 18:02:18	c*****a@gmail.com	CW	Matematika dan Ilmu Pengetahuan Alam	Pendidikan	4	4	5	4	5	4
1/6/2023 14:47:28	j*****4@gmail.com	JP	Psikologi dan Ilmu Sosial Budaya	Musik	5	4	5	4	5	5
1/5/2023 12:27:34	m*****7@gmail.com	M**y	Pendidikan	Pendidikan Anak Usia Dini	5	4	3	4	4	4
1/4/2023 17:40:56	e *****9@gmail.com	EP	Matematika dan Ilmu	Environmental Science	4	3	3	3	3	4

			Pengetahuan Alam							
1/6/2023 14:52:20	a *****6@ gmail.com	A**f	Teknologi Industri	Teknik Informatika	3	4	5	4	3	4
1/4/2023 13:42:31	s*****t@g mail.com	S***r	Bisnis dan Ekonomika	Reputation management	4	5	5	5	4	5
1/3/2023 17:26:01	a***** *****n@gmail.c om	SK	Psikologi dan Ilmu Sosial Budaya	Psikologi posistif	5	5	5	5	5	5
1/4/2023 10:04:20	b ***** ***5@gmail.com	T**a	Matematika dan Ilmu Pengetahuan Alam	Pendidikan	4	4	3	4	4	4
1/3/2023 20:07:32	b *****3@g mail.com	BC	Psikologi dan Ilmu Sosial Budaya	Bahasa Indonesia	4	3	4	4	4	4
1/3/2023 20:48:09	au*****n @gmail.com	F***h	Teknik Sipil dan Perencanaan	Hubungan AI dan children's literacy skill	5	5	4	5	5	5

1/6/2023 14:35:48	l*****i@gma il.com	L***** **i	Matematika dan Ilmu Pengetahuan Alam	Kimia	4	3	4	3	4	4
1/4/2023 11:53:16	d*****d@gmail.c om	P*****a	Matematika dan Ilmu Pengetahuan Alam	Machine Learning	4	4	3	4	4	4
1/5/2023 21:14:49	k*****h@g mail.com	KM	Bisnis dan Ekonomika	Administrasi Bisnis	4	3	4	4	4	4
1/4/2023 21:16:38	e *****3@ gmail.com	EP	Matematika dan Ilmu Pengetahuan Alam	Pendidikan Matematika	5	5	4	5	4	4
1/4/2023 22:14:35	p*****a@ gmail.com	N***** *****a	Bisnis dan Ekonomika	Manajemen Pendidikan	4	3	3	4	4	3
1/4/2023 12:57:28	n ***** *6@gmail.com	SA	CH	Pendidikan	5	5	5	5	5	5
1/3/2023 15:27:43	j*****n@g mail.com	JK	Teknologi Industri	Quantum mechanics, astrophysics	4	4	5	4	5	4

1/5/2023 21:01:54	e*****n@g mail.com	EA	Bisnis dan Ekonomika	State of the Art	5	5	5	5	5	5
1/4/2023 20:25:24	s*****@yah oo.com	I***** *****a	Olahraga dan Kesehatan	Pendidikan Jasmani Kesehatan dan Rekreasi	4	3	3	4	3	4
1/4/2023 22:46:17	p*****u@g mail.com	P***** **u	Psikologi dan Ilmu Sosial Budaya	Psikologi	4	4	5	5	4	4
1/4/2023 15:49:17	2*****1@students. uii.ac.id	F*****n	Teknologi Industri	Deep Learning & Image Processing	5	5	5	5	4	4
1/5/2023 15:33:00	c *****8@g mail.com	EM	Psikologi dan Ilmu Sosial Budaya	English	5	4	5	5	4	4
1/3/2023 18:57:49	t*****i@gmail .com	A***** ***i	Teologi	Teologi	4	4	3	5	3	4
1/5/2023 14:14:22	k*****i @gmail.com	KEA	Pendidikan	Pendidikan Guru Sekolah Dasar	4	4	4	5	4	5
1/3/2023 19:47:43	r *****9 @gmail.com	RR	Psikologi dan Ilmu Sosial Budaya	Sastra Perancis	5	5	5	5	4	4

1/5/2023 20:11:29	p*****a@gmail.com	P**u	Psikologi dan Ilmu Sosial Budaya	Cross culture in Indonesia	4	5	5	4	4	5
1/4/2023 22:15:45	j*****m@gmail.com	R*a	Manajemen	Science / Manajemen	5	5	4	5	4	5
1/5/2023 19:20:53	a*****a@gmail.com	AEP	SI	facial recognition	5	4	4	5	5	5
1/6/2023 18:38:50	a*****t@gmail.com	A***** ***** *****t	Fakultas	Arsitek	4	4	4	3	3	4
1/3/2023 20:26:42	o*****0@gmail.com	E*****n	Teknik	Teknik Elektro	5	4	5	3	4	4
1/5/2023 14:08:43	e*****3@gmail.com	ER	Matematika dan Ilmu Pengetahuan Alam	U-Net convolutional networks	4	3	3	4	4	4
1/6/2023 15:55:01	n***** *i@yahoo.com	NN	Matematika dan Ilmu Pengetahuan Alam	Teori aljabar	5	4	4	3	4	3

1/6/2023 18:49:08	a***y@cendekiaha rapan.sch.id	AV	Pendidikan	Pendidikan Anak Usia Dini	4	3	3	5	4	5
1/4/2023 19:37:18	d *****4 @civitas.ukrida.ac. id	DEC	Teknologi Industri	Sistem informasi	5	4	5	4	5	4
1/3/2023 13:07:25	k*****i@gma il.com	K***** *i	Kedokteran	Kebidanan	5	4	5	5	5	4
1/4/2023 17:36:06	f*****i@ gmail.com	W***n	Psikologi dan Ilmu Sosial Budaya	Filsafat	5	4	3	5	4	5
1/3/2023 11:09:25	n*****a@g mail.com	NN	Bisnis dan Ekonomika	Akuntansi Manajerial	4	5	5	5	4	4
1/4/2023 23:23:33	z*****1@ucl.ac.u k	ZC	Matematika dan Ilmu Pengetahuan Alam	Atomic and Molecular Physics	4	4	4	5	4	4
1/6/2023 10:59:59	s*****l@gmail.c om	SA	Matematika dan Ilmu Pengetahuan Alam	Fisika	5	4	4	4	4	4

1/4/2023 13:03:28	a*****n@gmail. com	SW	Pendidikan	Pendidikan Bahasa Inggris	4	5	3	4	4	3
1/4/2023 18:30:19	s*****e@ gmail.com	S*m	Sastra Inggris	Deep learning	4	5	4	4	4	4