JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 5, Mei 2025
E-ISSN : 3047-7824

Penerapan LSI untuk Topic Modelling Skripsi Sistem Informasi di Beberapa PTN Surabaya

Implementation of LSI for Topic Modeling of Information System Thesis in Several State Universities in Surabaya

Oryza Sativa Nufi[1*], Fitriya Mawadah Warahmah[2], Ailsa Aurellia[3], Mohammad Khusnu Milad[4]
[1,2,3,4]Sistem Informasi, Sains dan Teknologi, UIN Sunan Ampel Surabaya
Email: 09030622051@student.uinsby.ac.id[1], 09030622047@student.uinsby.ac.id[2], 09020622027@student.uinsby.ac.id[3], m.milad@uinsa.ac.id[4]*

Article Info

Article history :
Received : 08-05-2025
Revised : 10-05-2025
Accepted : 12-05-2025
Pulished : 14-05-2025

Abstract

This study applies Latent Semantic Indexing (LSI) for topic modelling on undergraduate thesis titles from Information Systems students at four public universities in Surabaya. The purpose of this research is to explore dominant research themes and academic trends that have emerged over recent years. A text mining approach is employed, where hundreds of thesis titles are collected from each university and pre-processed through tokenization, stopword removal, stemming, and term weighting using TF-IDF. The LSI method is then used to extract latent topics by reducing the dimensionality of the term-document matrix through singular value decomposition (SVD). The results indicate the presence of several dominant topics such as information system development, decision support systems, and data mining. These topics reflect the recurring areas of interest in the Information Systems curriculum across the universities studied. The study concludes that LSI is effective in identifying hidden semantic patterns and grouping thesis topics into meaningful clusters. These findings may support curriculum development and academic planning by highlighting students' thematic focus and institutional research directions.

Keywords* : *Information Systems, Latent Semantic Indexing, Topic Modelling

Abstrak

Penelitian ini menerapkan metode Latent Semantic Indexing (LSI) untuk melakukan pemodelan topik pada judul skripsi mahasiswa Program Studi Sistem Informasi dari empat perguruan tinggi negeri (PTN) di Surabaya. Tujuan dari penelitian ini adalah untuk mengidentifikasi tema-tema skripsi yang dominan dan menganalisis kecenderungan topik penelitian mahasiswa dalam beberapa tahun terakhir. Metodologi yang digunakan meliputi pengumpulan data berupa ratusan judul skripsi dari masing-masing PTN, kemudian dilakukan tahapan praproses teks seperti tokenisasi, penghapusan stopword, stemming, dan pembobotan menggunakan metode TF-IDF. Selanjutnya, metode LSI diterapkan dengan pendekatan dekomposisi nilai singular (SVD) untuk mengidentifikasi keterkaitan semantik antar istilah dan mereduksi dimensi data. Hasil pemodelan menunjukkan bahwa terdapat sejumlah topik dominan, antara lain pengembangan sistem informasi, sistem pendukung keputusan, dan data mining. Topik-topik tersebut mencerminkan fokus keilmuan yang berkembang dan menjadi tren dalam bidang Sistem Informasi di lingkungan perguruan tinggi yang diteliti. Berdasarkan hasil tersebut, penelitian ini menyimpulkan bahwa metode LSI efektif dalam mengelompokkan topik skripsi secara otomatis dan dapat menjadi alat bantu dalam evaluasi kurikulum serta perencanaan akademik.

Kata Kunci : Latent Semantic Indexing, Pemodelan Topik, Sistem Informasi

PENDAHULUAN

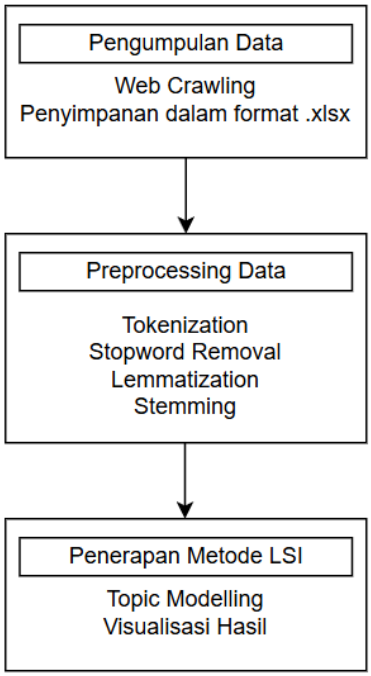
Dalam era digital, jumlah skripsi yang dihasilkan oleh mahasiswa di perguruan tinggi terus meningkat seiring dengan pertumbuhan Program Studi Sistem Informasi di berbagai institusi. Skripsi tidak hanya berperan sebagai syarat akademik untuk kelulusan, tetapi juga mencerminkan arah dan tren penelitian yang berkembang dalam lingkup keilmuan tertentu. Namun, volume dokumen yang besar ini dapat menjadi tantangan tersendiri, baik bagi mahasiswa yang mencari referensi, maupun bagi dosen yang ingin memetakan fokus riset mahasiswa secara umum. Oleh karena itu, dibutuhkan metode yang mampu membantu dalam memahami dan mengelompokkan topik-topik skripsi secara otomatis dan efisien.

Salah satu pendekatan yang dapat digunakan adalah *topic modelling*, yaitu teknik yang memungkinkan ekstraksi tema atau topik dari kumpulan teks dalam skala besar. (Alghamdi & Alfalqi, 2015) menyebutkan bahwa pemodelan topik merupakan cara efektif untuk menganalisis kumpulan data teks tidak terstruktur, dan Latent Semantic Indexing (LSI) menjadi salah satu metode awal yang terbukti berhasil dalam domain ini. LSI bekerja dengan memanfaatkan struktur semantik laten yang tersembunyi dalam dokumen, dan mampu mengidentifikasi kesamaan makna antara kata-kata yang mungkin tidak secara eksplisit terlihat.

(Kontostathis & Pottenger, 2006) menegaskan bahwa LSI berperan penting dalam mengurangi gangguan atau *noise* dalam data serta menyoroti dimensi semantik tersembunyi yang penting dalam klasifikasi dokumen. Di sisi lain, (Rani et al., 2017) menekankan bahwa LSI tidak hanya berguna untuk mereduksi dimensi data teks, tetapi juga efektif dalam menemukan hubungan semantik yang mendalam untuk mendukung pengelompokan dan pencarian dokumen yang lebih akurat. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan metode LSI dalam menganalisis judul dan abstrak skripsi mahasiswa Sistem Informasi di beberapa perguruan tinggi negeri Surabaya, guna mengidentifikasi tren topik dominan dan memberikan kontribusi bagi pengembangan akademik di tingkat program studi.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis analisis teks untuk memetakan topik-topik skripsi mahasiswa Program Studi Sistem Informasi dari beberapa perguruan tinggi negeri di Surabaya. Data diambil dari repositori skripsi secara daring, kemudian diproses melalui tahapan *web crawling*, praproses teks (tokenisasi, penghapusan *stopword*, *stemming*, dan *lemmatization*), serta pembobotan menggunakan *Bag of Word*. Selanjutnya, metode *Latent Semantic Indexing* (LSI) diterapkan untuk mengelompokkan dokumen berdasarkan kemiripan semantik. Menurut (Guntoorkar & Pearce, n.d.) LSI membentuk matriks term-dokumen dan menerapkan dekomposisi untuk menurunkan dimensi, sehingga dapat mengungkap hubungan semantik laten antar dokumen.



Gambar 1. Metode Penelitian

Pengumpulan Data

Data dikumpulkan menggunakan teknik *web crawling* dari repositori skripsi milik empat perguruan tinggi negeri di Surabaya. Skrip *crawler* ditulis dalam bahasa *Python* dengan *library BeautifulSoup* untuk mengekstrak elemen HTML yang berisi informasi judul, abstrak, penulis, tahun, dan link dokumen. Data yang berhasil dikumpulkan kemudian disimpan dalam format Excel (.xlsx) untuk diproses pada tahap selanjutnya.

Preprocessing Data

Tahap ini dilakukan untuk membersihkan dan menyiapkan data teks agar dapat dianalisis secara efektif oleh algoritma LSI. Proses ini mencakup beberapa langkah:

- Tokenization: memecah teks menjadi unit kata atau istilah individual.
- Stopword Removal: menghapus kata-kata umum (seperti “yang”, “dan”, “adalah”) yang tidak membawa makna penting dalam analisis.
- Lemmatization: mengubah kata menjadi bentuk dasarnya dengan mempertimbangkan konteks linguistik.
- Stemming: menghapus imbuhan untuk mendapatkan bentuk akar kata tanpa mempertimbangkan makna.

Langkah-langkah ini dilakukan secara berurutan menggunakan *library* NLP dalam Python seperti NLTK dan Sastrawi.

Penerapan Metode LSI

Setelah proses praproses selesai, data dibobot menggunakan metode TF-IDF untuk membentuk matriks kata-dokumen. Selanjutnya, metode Latent Semantic Indexing (LSI) diterapkan untuk mengidentifikasi keterkaitan semantik antar dokumen dan mengelompokkan skripsi berdasarkan topik. Hasil dari penerapan LSI divisualisasikan dalam bentuk grafik koherensi

topik, distribusi topik berdasarkan tahun, serta *wordcloud* untuk masing-masing topik guna mempermudah interpretasi hasil.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari penerapan metode *Latent Semantic Indexing* (LSI) terhadap kumpulan judul dan abstrak skripsi mahasiswa Program Studi Sistem Informasi di empat perguruan tinggi negeri di Surabaya. Hasil yang ditampilkan meliputi proses pengumpulan data, tahap praproses teks, pemodelan topik, serta visualisasi yang menggambarkan distribusi topik berdasarkan tahun. Pembahasan dilakukan dengan memberikan interpretasi terhadap hasil yang diperoleh serta mengaitkannya dengan teori dan penelitian terdahulu untuk memperkuat pemaknaan. Melalui tahapan ini, diperoleh gambaran tematik yang dapat menjadi dasar dalam memahami tren penelitian mahasiswa serta arah pengembangan keilmuan di bidang Sistem Informasi.

Hasil Pengumpulan Data

Sebelum dilakukan analisis lebih lanjut, langkah awal dalam penelitian ini adalah menghitung jumlah skripsi yang berhasil dikumpulkan berdasarkan tahun terbitnya. Informasi ini memberikan gambaran awal tentang sebaran data yang digunakan dalam pemodelan topik, serta menunjukkan tren produktivitas penelitian mahasiswa dari tahun ke tahun. Data jumlah skripsi dari tahun 2018 hingga 2024 pada empat Perguruan Tinggi Negeri di Surabaya disajikan dalam Tabel 1 berikut.

Tabel 1. Pengumpulan Data

Tahun	Jumlah Skripsi
2018	94
2019	100
2020	96
2021	89
2022	119
2023	182
2024	218

Preprocessing Data

Setelah data skripsi berhasil dikumpulkan dari keempat perguruan tinggi, dilakukan *preprocessing* untuk membersihkan dan menyiapkan data teks sebelum dianalisis lebih lanjut. Kami akan melakukan *preprocessing* pada salah satu kolom "Judul" dan "Abstrak" sesuai dengan Hasil Pengumpulan Data sebelumnya. Tabel 2 berikut menyajikan perbandingan jumlah data sebelum dan sesudah dilakukan *preprocessing*.

Tabel 2. Preprocessing Data

	Sebelum Preprocessing	Setelah Preprocessing
Judul	Rancangan master plan sistem teknologi informasi pada dinas komunikasi dan informatika kabupaten Nganjuk menggunakan metode Ward and Peppard.	[rancang, master, plan, sistem, teknologi, informasi, dina, komunikasi, informatika, kabupaten, nganjuk, guna, metode, ward, peppard]
Abstrak	Dalam Tugas Pokoknya dinas kominfo membantu bupati dalam melaksanakan sebagian urusan rumah tangga daerah dibidang komunikasi dan informatika. Tentunya diperlukan dukungan teknologi informasi dan sistem informasi yang handal. Untuk memastikan penggunaakn SI/TI tersebut mendukung pemerintah, maka diperlukan masterplan SI/TI atau rancangan strategi yang terkait dengan SI/TI yang berbentuk dokumen disebut dengan masterplan TI. Ada beberapa metode dalam membuat rancangan master plan Teknologi Informasi. Pada penelitian ini menggunakan metode Ward And Peppard. Metode ini terdiri dari tahapan masukan dan keluaran.	[tugas, pokok, dinas, kominfo, bantu, bupati, laksana, urus, rumah, tangga, daerah, bidang, komunikasi, informatika, perlu, dukung, teknologi, informasi, sistem, informasi, handal, pasti, guna, si, ti, dukung, pemerintah, perlu, masterplan, si, ti, rancang, strategi, kait, si, ti, bentuk, dokumen, sebut, masterplan, ti, ada, metode, buat, rancang, master, plan, teknologi, informasi, teliti, guna, metode, ward, peppard, metode, terdiri, tahap, masuk, keluar]

Hasil dari *preprocessing data* berupa *term* atau matriks yang berisi kata-kata yang muncul berulang-ulang. Kami menggunakan model *Bag of words* untuk menghitung jumlah kemunculan setiap kata pada term atau matriks kata-kata tersebut. Hasil perhitungan jumlah setiap kata tersebut digunakan dalam dalam perhitungan distribusi yang ada pada LSI.

Tabel 3. Frekuensi Kemunculan Kata

No	Kata	Jumlah Kemunculan
1.	data	1410
2.	sistem	1247
3.	informasi	860
4.	metode	853
5.	system	833

Berdasarkan hasil analisis frekuensi kemunculan kata, lima kata yang paling sering muncul dalam seluruh dokumen adalah *data*, *sistem*, *informasi*, *metode*, dan *system*. Kata-kata tersebut menunjukkan bahwa sebagian besar skripsi memiliki fokus pada pengolahan data, perancangan sistem, serta penggunaan metode tertentu dalam pengembangan solusi teknologi informasi. Dominasi kata-kata tersebut juga mencerminkan karakteristik umum dari penelitian di bidang Sistem Informasi, yang umumnya berkaitan dengan penerapan teknologi untuk menyelesaikan masalah organisasi atau bisnis.

LSI Topic Modelling

Setelah proses pemodelan dilakukan dengan metode *Latent Semantic Indexing* (LSI), diperoleh sejumlah topik laten yang terbentuk dari kumpulan kata dengan bobot tertinggi dalam setiap topik. Setiap topik mencerminkan representasi semantik yang tersembunyi di balik dokumen-dokumen skripsi, yang diinterpretasikan berdasarkan kombinasi kata dominannya. Tabel berikut menyajikan sepuluh topik yang dihasilkan dari pemodelan LSI beserta sepuluh kata utama yang memiliki bobot tertinggi dalam masing-masing topik.

Tabel 4. Hasil Latent Semantic Indexing (LSI)

Hasil LSI
(0, '-0.408*"sistem" + -0.273*"informasi" + -0.214*"metode" + -0.207*"system" + -0.196*"aplikasi" + -0.176*"information" + -0.173*"user" + -0.170*"model" + -0.163*"pengguna" + -0.148*"teknologi"')
(1, '-0.432*"system" + -0.362*"information" + 0.356*"sistem" + -0.227*"user" + -0.181*"application" + 0.174*"informasi" + 0.155*"metode" + -0.123*"service" + -0.120*"method" + -0.119*"quality"')
(2, '-0.392*"sistem" + 0.386*"the" + 0.299*"of" + 0.243*"digital" + -0.202*"system" + 0.180*"aplikasi" + 0.176*"in" + 0.173*"model" + 0.173*"to" + -0.164*"information"')
(3, '-0.387*"the" + 0.320*"aplikasi" + -0.239*"sistem" + 0.222*"pengguna" + -0.179*"of" + 0.178*"perceived" + -0.169*"in" + -0.158*"digital" + 0.158*"variabel" + 0.157*"user"')
(4, '0.394*"sistem" + -0.349*"teknologi" + -0.256*"informasi" + -0.218*"prose" + 0.177*"user" + 0.144*"application" + -0.140*"dasar" + -0.139*"information" + -0.139*"bisnis" + -0.132*"kerja"')
(5, '0.444*"website" + -0.293*"perceived" + 0.272*"aplikasi" + 0.203*"usability" + -0.172*"sistem" + -0.171*"model" + 0.166*"metode" + -0.165*"informasi" + -0.159*"teknologi" + -0.143*"intention"')
(6, '-0.404*"user" + 0.376*"system" + 0.318*"aplikasi" + -0.307*"application" + -0.243*"website" + 0.242*"information" + 0.153*"pengguna" + -0.149*"satisfaction" + -0.138*"informasi" + -0.136*"variable"')
(7, '-0.519*"website" + 0.411*"aplikasi" + 0.325*"application" + -0.156*"surabaya" + -0.139*"variabel" + 0.118*"user" + -0.107*"system" + -0.105*"quality" + -0.101*"kualitas" + 0.100*"mobile"')
(8, '0.309*"system" + 0.244*"siswa" + -0.244*"level" + 0.233*"dasar" + 0.207*"program" + -0.202*"informasi" + 0.174*"teknologi" + 0.166*"sekolah" + 0.159*"pembelajaran" + 0.143*"pendidikan"')
(9, '0.336*"aplikasi" + 0.335*"informasi" + -0.309*"nilai" + -0.286*"metode" + -0.238*"model" + -0.148*"prose" + 0.148*"system" + -0.132*"level" + 0.130*"digital" + 0.117*"user"')

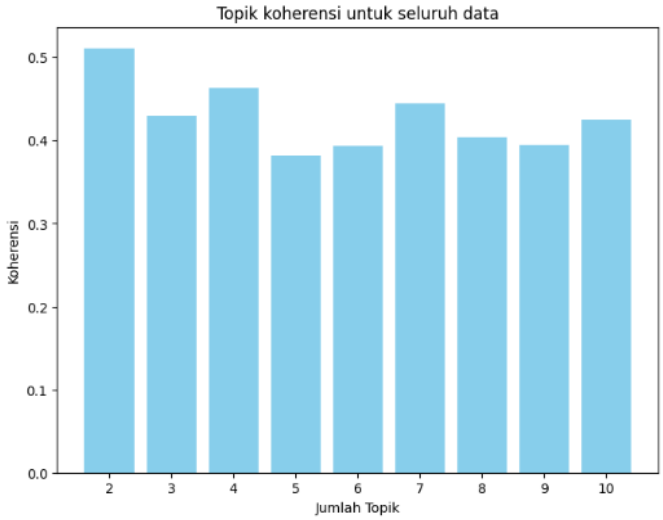
Berdasarkan hasil pada tabel, terlihat bahwa beberapa kata seperti *sistem*, *informasi*, *aplikasi*, *user*, dan *website* muncul secara konsisten di berbagai topik. Hal ini menunjukkan bahwa fokus penelitian mahasiswa Prodi Sistem Informasi banyak berkisar pada pengembangan sistem informasi, teknologi aplikasi berbasis web, serta pengalaman pengguna. Selain itu, munculnya kata

seperti *pendidikan*, *siswa*, dan *pembelajaran* pada salah satu topik menunjukkan adanya keterkaitan penelitian dengan dunia pendidikan. Keberagaman kata dalam tiap topik memperlihatkan luasnya cakupan riset dan minat mahasiswa selama periode yang dianalisis.

Visualisasi Hasil

A. TOP koherensi untuk seluruh data

Pada proses *topic coherence*, fungsi *CoherenceModel* digunakan untuk menghitung skor koherensi topik, yang kemudian digunakan untuk menentukan jumlah topik yang paling sesuai untuk diterapkan pada model LSI. Skor koherensi tertinggi menunjukkan jumlah topik yang optimal untuk model tersebut. Dalam penelitian ini, pencarian jumlah topik dilakukan dengan menggunakan data gabungan dari seluruh korpus, mencakup jumlah topik mulai dari 2 hingga 10. Berdasarkan hasil metode ini, jumlah topik yang memberikan nilai koherensi tertinggi adalah 6, sehingga dianggap sebagai jumlah topik yang paling sesuai.

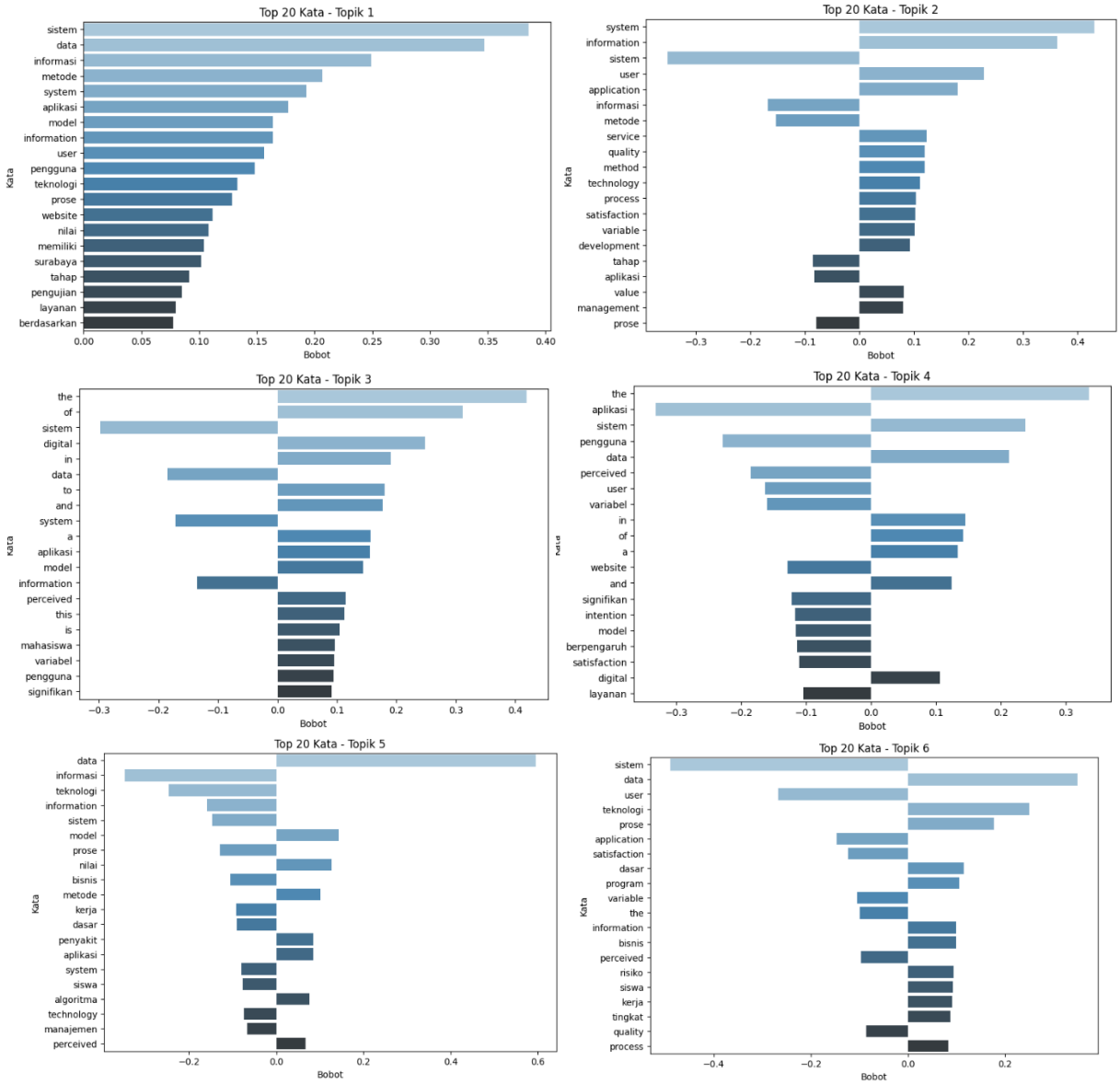


Gambar 2. Topik Koherensi

B. TOP 20 kata

Grafik di atas menunjukkan *Top 20 Kata* dalam Topik 1 berdasarkan hasil pemodelan topik menggunakan pendekatan LSI. Kata dengan bobot tertinggi seperti *sistem*, *data*, dan *informasi* menunjukkan bahwa topik ini berfokus pada tema yang berkaitan dengan sistem informasi, teknologi, dan pengolahan data. Kata-kata lain seperti *metode*, *aplikasi*, *pengguna*, dan *teknologi* menguatkan indikasi bahwa topik ini menyoroti aspek teknis dan implementasi suatu sistem, termasuk pengujian, layanan, dan evaluasi metode tertentu. Visualisasi ini membantu memahami konteks utama dari topik yang dianalisis, yang tampaknya relevan dengan penelitian terkait teknologi dan informasi.

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 5, Mei 2025
E-ISSN : 3047-7824

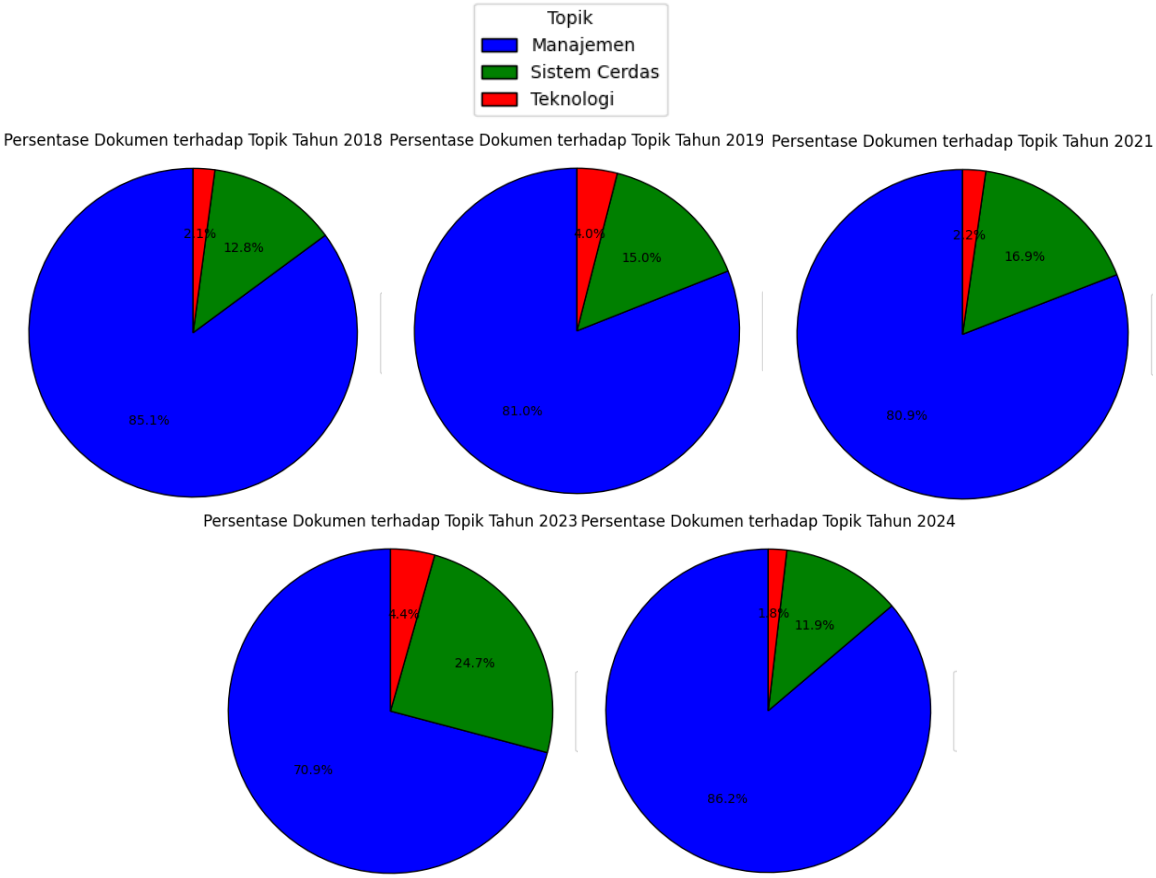


Gambar 3. Top 20 Kata

C. Presentase dokumen terhadap topik tahun 2018 – 2024

Gambar Presentase Dokumen terhadap Topik Tahun 2018–2024 menggambarkan distribusi proporsi dokumen berdasarkan topik utama untuk setiap tahun, yang mencerminkan tren penelitian di Prodi Sistem Informasi. Topik-topik seperti keamanan siber dan pengembangan aplikasi tampak mendominasi pada tahun-tahun tertentu, mencerminkan respons terhadap perkembangan teknologi dan kebutuhan industri.

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 5, Mei 2025
E-ISSN : 3047-7824



Perubahan proporsi dari tahun ke tahun menunjukkan adanya pergeseran fokus penelitian mahasiswa yang dapat dipengaruhi oleh dinamika teknologi, kebijakan akademik, maupun kebutuhan pasar. Visualisasi ini tidak hanya membantu akademisi dalam menyesuaikan kurikulum agar tetap relevan, tetapi juga menyoroti area penelitian yang kurang terwakili, sehingga dapat mendorong eksplorasi topik baru yang strategis dan mendukung arah pengembangan akademik program studi.

KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan Latent Semantic Indexing (LSI) berhasil mengelompokkan dan menganalisis tema skripsi mahasiswa Prodi Sistem Informasi di empat Perguruan Tinggi Negeri Surabaya. Hasil analisis LSI memberikan wawasan tentang tren penelitian dari tahun 2018 hingga 2024, dengan topik populer seperti pengembangan aplikasi dan keamanan siber. Teknik ini mempermudah mahasiswa dalam menemukan referensi yang relevan melalui pengelompokan skripsi berdasarkan tema, sehingga meningkatkan efisiensi dalam menentukan topik penelitian.

Tahapan preprocessing data yang dilakukan, seperti tokenisasi, penghapusan stopwords, stemming, dan lemmatization, terbukti optimal dalam mempersiapkan data untuk analisis. Proses ini menghasilkan model topik yang jelas dan dapat digunakan untuk memahami hubungan semantik antar dokumen. Visualisasi seperti koherensi topik, wordcloud, dan distribusi dokumen per tahun mendukung pemahaman hasil analisis.

Penelitian ini juga memberikan manfaat strategis bagi dosen dalam merancang kurikulum yang relevan dengan tren industri dan teknologi terkini. Selain itu, pendekatan ini berkontribusi

pada pengembangan ilmu di bidang sistem informasi, terutama dalam analisis teks dan pemodelan topik. Penelitian ini membuka peluang eksplorasi lebih lanjut terhadap metode lain yang dapat meningkatkan akurasi dan efisiensi analisis skripsi di masa depan.

DAFTAR PUSTAKA

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In *Mining Text Data* (Vol. 9781461432). https://doi.org/10.1007/978-1-4614-3223-4

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1), 147–153. https://doi.org/10.14569/ijacsa.2015.060121

Dayera, Musa Bundaris Palungan, F. O. (2024). G-Tech : Jurnal Teknologi Terapan. *G-Tech : Jurnal Teknologi Terapan*, *8*(1), 186–195. https://ejournal.uniramalang.ac.id/index.php/g-tech/article/view/1823/1229

Fernando, E. H., & Toba, H. (2020). Pemanfaatan Latent Semantic Indexing untuk Mengukur Potensi Kerjasama Jurnal Ilmiah Lintas Universitas. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(3), 489–503. https://doi.org/10.28932/jutisi.v6i3.2894

Guntoorkar, P., & Pearce, C. (n.d.). *Latent Semantic Indexing : A Regularized approach to large-scale modeling*.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing and Management*, *42*(1 SPEC. ISS), 56–73. https://doi.org/10.1016/j.ipm.2004.11.007

Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2021). Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, *117*(May 2020), 102096. https://doi.org/10.1016/j.artmed.2021.102096

Rani, M., Dhar, A. K., & Vyas, O. P. (2017). Semi-automatic terminology ontology learning based on topic modeling. *Engineering Applications of Artificial Intelligence*, *63*, 108–125. https://doi.org/10.1016/j.engappai.2017.05.006

Reynolds, T. P., & Mesbahi, M. (2020). The Crawling Phenomenon in Sequential Convex Programming. *Proceedings of the American Control Conference*, *2020-July*, 3613–3618. https://doi.org/10.23919/ACC45564.2020.9147550

Wang, Q., Xu, J., Li, H., & Craswell, N. (2013). Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems*, *31*(1). https://doi.org/10.1145/2414782.2414787