

**PEMETAAN TOPIK SKRIPSI MAHASISWA BERBASIS *TOPIC MINING*
MENGUNAKAN *LATENT DIRICHLET ALLOCATION(LDA)***

SKRIPSI

Oleh :

AHMAD ARSHA ALBAR
NIM. 210605110055



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

**PEMETAAN TOPIK SKRIPSI MAHASISWA BERBASIS *TOPIC MINING*
MENGUNAKAN *LATENT DIRICHLET ALLOCATION(LDA)***

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
AHMAD ARSHA ALBAR
NIM. 210605110055

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

HALAMAN PERSETUJUAN

PEMETAAN TOPIK SKRIPSI MAHASISWA BERBASIS *TOPIC MINING* MENGUNAKAN *LATENT DIRICHLET ALLOCATION (LDA)*

SKRIPSI

Oleh :
AHMAD ARSHA ALBAR
NIM. 210605110055

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 11 Mei 2026

Pembimbing I,



Dr. M. Ainul Yaqin, M.Kom
NIP. 19761013 200604 1 004

Pembimbing II,



Ashri Shabrina Afrah, M.T
NIP. 19900430 202012 2 003

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M.Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

PEMETAAN TOPIK SKRIPSI MAHASISWA BERBASIS *TOPIC MINING* MENGUNAKAN *LATENT DIRICHLET ALLOCATION(LDA)*

SKRIPSI

Oleh :
AHMAD ARSHA ALBAR
NIM. 210605110055

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 11 Mei 2026

Susunan Dewan Penguji

Ketua Penguji : Prof. Dr. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

Anggota Penguji I : Dr. Yunifa Miftachul Arif, M.T
NIP. 19830616 201101 1 004

Anggota Penguji II : Dr. M. Ainul Yaqin, M.Kom
NIP. 19761013 200604 1 004

Anggota Penguji III : Ashri Shabrina Afrah, M.T
NIP. 19900430 202012 2 003

()
()
()
()

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M.Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Ahmad Arsha Albar
NIM : 210605110055
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Pemetaan Topik Skripsi Mahasiswa Berbasis *Topic Mining* Menggunakan *Latent Dirichlet Allocation (LDA)*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 11 Mei 2026
Yang membuat pernyataan,



Ahmad Arsha Albar
NIM. 210605110055

MOTTO

... Motto hari ini adalah tanpa moto sama sekali ...

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur ke hadirat Allah SWT atas segala rahmat, kemudahan, dan kekuatan-Nya, skripsi ini dapat terselesaikan dengan baik. Untuk itu, penulis persembahkan skripsi ini kepada;

1. Kedua orang tua saya yang selalu memberikan dukungan baik moril maupun materi, serta kasih sayang dan doa yang tidak pernah terputus dari awal penulis lahir hingga saat ini, terima kasih telah menjadi orang tua yang sempurna, menjadi teladan dalam hidupku. Terimakasih atas perhatian, dukungan dan doa yang tak pernah terputus.
2. Dosen penulis, yakni Bapak Dr M.Ainul yaqin, M.Kom., dan Ibu Ashri Shabrina Afrah, M.T yang telah banyak membimbing dan memberikan motivasi kepada penulis baik dalam perkuliahan serta dalam proses penulisan skripsi ini.
3. Partner terbaik penulis, Rizqi Amalia Kartika, yang kehadirannya menjadi teduh di tengah lelah, serta menjadi sandaran diantara keraguan, menjadi arah ketika langkah mulai goyah. Terimakasih senantiasa mengiringi penulis hingga skripsi ini dapat terselesaikan.

Serta semua pihak yang telah membantu dan memberikan dukungan dalam bentuk apapun, yang tidak dapat disebutkan satu per satu.

KATA PENGANTAR

Puji Syukur atas kehadiran Allah Subhanahu wa ta'ala, karena berkat rahmat dan petunjuk-Nya penulis dapat menyelesaikan skripsi yang berjudul “Pemetaan Topik Skripsi Mahasiswa Berbasis *Topic Mining* Menggunakan *Latent Dirichlet Allocation*(LDA).” Shalawat serta salam selalu tercurahkan kepada Nabi Muhammad Shallallahu ‘alaihi wasallam, yang telah membawa kita dari zaman kegelapan menuju jalan yang terang benderang yaitu Agama Islam.

Selama penulisan skripsi ini, penulis menyadari bahwa banyak pihak yang terlibat dalam membantu, membimbing, menyemangati, dan mendoakan penulis hingga berada di titik ini. Dukungan moril maupun materil yang telah penulis terima tidak akan pernah penulis lupakan dan semoga akan dibalas dengan berlipat lipat kebaikan. Untuk itu, penulis ingin mengucapkan rasa terima kasih kepada:

1. Prof. Dr. Hj. Ilfi Nurdiana, M.Si., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Supriyono, M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. M. Ainul Yaqin, M.Kom selaku Dosen Pembimbing I dan Ashri Shabrina Afrah, M.T selaku dosen pembimbing II yang telah membimbing dan menuntun penulis dalam menyelesaikan skripsi ini.

5. Prof. Dr. Muhammad Faisal, M.T selaku Ketua Penguji yang telah menguji dan memberikan banyak saran untuk menyelesaikan skripsi ini.
6. Dr. Yunifa Miftachul Arif, M.T selaku dosen penguji II yang telah menguji serta memberikan masukan dalam penulisan skripsi ini.
7. Seluruh Dosen dan Jajaran Staf Program Studi Teknik Informatika yang telah memberikan banyak bantuan dalam menyelesaikan skripsi ini.

kekurangan. Penulis berharap skripsi ini dapat diterima dengan baik bagi masyarakat dan memberikan kontribusi terhadap ilmu pengetahuan. Semoga Allah Subhanahu wa ta'ala menerima hasil karya saya sebagai ibadah dan meridoi langkah saya selanjutnya.

DAFTAR ISI

HALAMAN PENGAJUAN.....	ii
HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN.....	iv
PERNYATAAN KEASLIAN TULISAN.....	v
MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xiii
DAFTAR TABEL	xiv
ABSTRAK	xv
ABSTRACT	xvi
مستخلص البحث.....	xvii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Batasan Masalah.....	5
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
BAB II STUDI PUSTAKA.....	7
2.1 Penelitian Terdahulu	7
2.2 Skripsi Mahasiswa	24
2.3 <i>Topic mining</i>	17
2.4 <i>Latent Dirichlet Allocation (LDA)</i>	19
2.5 <i>Text Processing</i>	23
2.5.1 <i>N-gram</i>	24
2.5.2 TF-IDF	26
2.5.3 <i>Coherence Score</i>	27
2.6 <i>Python</i>	28
BAB III METODE PENELITIAN	30
3.1 Desain Penelitian	30
3.1.1 Desain Sistem.....	31
3.2 Data	33

3.3 <i>Text Processing</i>	34
3.4 Membangun <i>Corpus</i>	36
3.5 Implementasi Metode	42
3.6 Skenario Uji Coba	47
3.7 Evaluasi	50
BAB IV HASIL DAN PEMBAHASAN	53
4.1 Data Penelitian	53
4.1.1 Seleksi Data	56
4.2 <i>Text Processing</i>	60
4.2.1 <i>Case Folding</i> dan <i>Cleaning</i>	61
4.2.2 <i>Tokenizing</i>	62
4.2.3 <i>Stopword Removal</i>	63
4.2.4 <i>Stemming</i>	63
4.3 Membangun <i>Corpus</i>	64
4.3.1 Bigram	65
4.3.2 <i>Dictionary</i>	67
4.3.3 <i>Filtering</i> Kata	68
4.3.4 <i>Corpus</i>	69
4.4 Pemodelan LDA	70
4.4.1 Proses Pemodelan LDA	70
4.4.2 Optimasi Jumlah Topik	71
4.4.3 Hasil Topik LDA	74
4.5 Evaluasi Model	74
4.5.1 Evaluasi Berdasarkan <i>Coherence Score</i>	75
4.5.2 Evaluasi Berdasarkan Distribusi Topik	76
4.5.3 Analisis Kualitas Model	77
4.6 Hasil <i>Topic mining</i>	78
4.6.1 Visualisasi Distribusi Topik (pyLDAvis)	78
4.6.2 Visualisasi Tren Topik Penelitian	81
4.6.3 Visualisasi <i>Wordcloud</i> Topik	84
4.7 Pembahasan	87
4.7.1 Analisis Dominasi Topik Penelitian	87
4.7.2 Analisis Distribusi dan Ketimpangan Topik	88
4.7.3 Analisis Hubungan Antar Topik	89
4.7.4 Analisis Kualitas Model	90

4.7.5 Analisis Kelebihan dan Keterbatasan Model	91
4.7.7 Implikasi Penelitian	92
4.8 Integrasi Sains dan Islam	93
BAB V KESIMPULAN DAN SARAN	100
5.1 Kesimpulan	100
5.2 Saran	101
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 2. 1 Ilustrasi <i>Latent Dirichlet Allocation</i>	20
Gambar 2. 2 Plat Notasi LDA.....	20
Gambar 3. 1 Desain Penelitian.....	30
Gambar 3. 2 Desain Sistem.....	32
Gambar 4. 1 Optimasi Jumlah Topik.....	72
Gambar 4. 2 Visualisasi pyLDAvis	78
Gambar 4. 3 Distribusi Topik Bidang TI.....	81
Gambar 4. 4 <i>Wordcloud</i> Topik 1	85
Gambar 4. 5 <i>Wordcloud</i> Topik 2	85
Gambar 4. 6 <i>Wordcloud</i> Topik 6	85
Gambar 4. 7 <i>Wordcloud</i> Topik 9	86

DAFTAR TABEL

Tabel 2. 1 Penelitian terkait.....	12
Tabel 3. 1 Dataset.....	27
Tabel 3. 2 Proses <i>Tokenizing</i>	28
Tabel 3. 3 Proses <i>Stopword</i>	28
Tabel 3. 4 Proses <i>Stemming</i>	29
Tabel 3. 5 <i>Term</i> Frequency (TF).....	31
Tabel 3. 6 Normalisasi <i>Term Frequency</i> (TF)	31
Tabel 3. 7 TF-IDF	33
Tabel 3. 8 Insialisasi Topik.....	35
Tabel 4. 1 Informasi Dataset.....	54
Tabel 4. 2 Nama Kolom Dataset.....	54
Tabel 4. 3 Distribusi Data Berdasarkan Tahun.....	55
Tabel 4. 4 Statistik Panjang Abstrak.....	55
Tabel 4. 5 Contoh Data Awal.....	55
Tabel 4. 6 Rincian Seleksi Data	58
Tabel 4. 7 Rincian Alasan Data Gugur	59
Tabel 4. 8 <i>Case Folding</i> dan <i>Cleaning</i>	61
Tabel 4. 9 <i>Tokenizing</i>	62
Tabel 4. 10 <i>Stopword Removal</i>	63
Tabel 4. 11 <i>Stemming</i>	63
Tabel 4. 12 Hasil <i>Bigram</i>	66
Tabel 4. 13 Contoh <i>Dictionary</i>	67
Tabel 4. 14 Contoh <i>Filtering</i> Kata.....	68
Tabel 4. 15 Contoh <i>Corpus</i>	69
Tabel 4. 16 Hasil Topik LDA	74
Tabel 4. 17 Nilai <i>Coherence</i> Berdasarkan Jumlah Topik	75
Tabel 4. 18 Distribusi Proporsi Topik.....	76

ABSTRAK

Albar, Ahmad Arsha. 2026. **Pemetaan Tren Topik Skripsi Mahasiswa Berbasis Topic Mining Menggunakan Latent Dirichlet Allocation(Lda)**. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. M. Ainul Yaqin, M.Kom (II) Ashri Shabrina Afrah, M.T.

Kata kunci: *Topic Mining, Latent Dirichlet Allocation (LDA), Text Mining, Coherence Score.*

Pemilihan topik skripsi merupakan salah satu tahapan penting dalam penelitian mahasiswa, namun seringkali menjadi kendala karena keterbatasan referensi dan kecenderungan topik yang berulang. Kondisi tersebut dapat menyebabkan rendahnya variasi dan inovasi topik penelitian di lingkungan akademik. Oleh karena itu, diperlukan pendekatan yang mampu mengidentifikasi dan memetakan tren topik skripsi secara otomatis berdasarkan data skripsi yang telah tersedia. Penelitian ini bertujuan untuk mengidentifikasi dan memetakan tren topik skripsi mahasiswa menggunakan metode *topic mining* berbasis *Latent Dirichlet Allocation* (LDA). Data yang digunakan diperoleh dari *repository e-theses* UIN Maulana Malik Ibrahim Malang pada rentang tahun 2020–2025 dengan jumlah awal 3961 dokumen. Setelah melalui filter awal domain Teknik Informatika dan seleksi token TI, sebanyak 3349 dokumen atau 84,55% digunakan sebagai corpus final LDA, sedangkan 612 dokumen gugur karena tidak memenuhi kriteria seleksi. Pemodelan dilakukan menggunakan LDA dengan jumlah final 9 topik dan memperoleh nilai *coherence score* sebesar 0,429723. Hasil pemetaan bidang menunjukkan bahwa 2090 dokumen berhasil diklasifikasikan ke dalam enam bidang Teknik Informatika. Bidang paling dominan adalah *Software Engineering* sebanyak 583 dokumen atau 27,89%, diikuti oleh *Intelligence System* sebanyak 579 dokumen atau 27,70%, *IoT & Robotic* sebanyak 359 dokumen atau 17,18%, *Multimedia & Game* sebanyak 281 dokumen atau 13,44%, Sistem Informasi sebanyak 183 dokumen atau 8,76%, dan *Web & Mobile Programming* sebanyak 105 dokumen atau 5,02%. Hasil penelitian menunjukkan bahwa LDA mampu memberikan gambaran tren topik skripsi mahasiswa secara terstruktur dan dapat digunakan sebagai referensi dalam pemilihan topik skripsi yang lebih relevan.

ABSTRACT

Albar, Ahmad Arsha. 2026. **Mapping Student Thesis Topic Trends Based on Topic Mining Using Latent Dirichlet Allocation (LDA)**. Undergraduate Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisors: (I) Dr. M. Ainul Yaqin, M.Kom. (II) Ashri Shabrina Afrah, M.T.

Keywords: Topic Mining, Latent Dirichlet Allocation (LDA), Text Mining, Coherence Score.

The selection of a thesis topic is an important stage in student research, yet it often becomes a challenge due to limited references and the tendency of recurring topics. This condition may lead to low variation and innovation in academic research topics. Therefore, an approach is needed to automatically identify and map thesis topic trends based on existing thesis data. This study aims to identify and map student thesis topic trends using a topic mining method based on Latent Dirichlet Allocation (LDA). The data were obtained from the e-theses repository of UIN Maulana Malik Ibrahim Malang from 2020 to 2025, with an initial total of 3,961 documents. After applying the Informatics Engineering domain filter and TI token selection, 3,349 documents, or 84.55%, were used as the final LDA corpus, while 612 documents were excluded because they did not meet the selection criteria. The modeling process was conducted using LDA with a final number of 9 topics and obtained a coherence score of 0.429723. The field mapping results show that 2,090 documents were successfully classified into six Informatics Engineering fields. The most dominant field was Software Engineering with 583 documents, or 27.89%, followed by Intelligence System with 579 documents, or 27.70%, IoT & Robotic with 359 documents, or 17.18%, Multimedia & Game with 281 documents, or 13.44%, Information System with 183 documents, or 8.76%, and Web & Mobile Programming with 105 documents, or 5.02%. The results show that LDA is able to provide a structured overview of student thesis topic trends and can be used as a reference in selecting more relevant thesis topics.

مستخلص البحث

ألبار، أحمد عرشا. ٢٠٢٦. رسم خريطة لاتجاهات موضوعات رسائل الطلاب بناءً على تنقيب الموضوعات باستخدام التخصيص الديرشلييه الكامن (LDA). رسالة بكالوريوس. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مالك إبراهيم الإسلامية الحكومية مالانج. المشرفون: (I) د. م. عين اليقين، م. كوم. (II) أشري شابرينا أفراح، م. ت.

الكلمات المفتاحية: تنقيب الموضوعات، التخصيص الديرشلييه الكامن (LDA)، تنقيب النصوص، درجة الاتساق.

يُعدُّ اختيار موضوع رسالة التخرج من المراحل المهمة في البحث العلمي لدى الطلاب، إلّا أنّه غالبًا ما يُواجه بعقبات ناجمة عن محدودية المراجع وتكرار الموضوعات المختارة، ممّا يؤدي إلى انخفاض التنوع والابتكار في موضوعات الأبحاث داخل البيئة الأكاديمية. لذلك، تبرز الحاجة إلى منهجية قادرة على تحديد اتجاهات موضوعات الرسائل الجامعية وتصنيفها بصورة آلية اعتمادًا على البيانات المتوفرة مسبقًا. تهدف هذه الدراسة إلى تحديد وتحليل اتجاهات موضوعات رسائل التخرج لدى الطلاب باستخدام أسلوب تنقيب الموضوعات القائم على نموذج «Latent Dirichlet Allocation (LDA)» تمّ الحصول على البيانات من مستودع الرسائل الإلكترونية بجامعة UIN Maulana Malik Ibrahim Malang خلال الفترة الممتدة من ٢٠٢٠ إلى ٢٠٢٥، بعددٍ أولي بلغ ٣٩٦١ وثيقة. وبعد إجراء التصنيفية الأولية في مجال هندسة المعلوماتية واختيار الرموز الخاصة بتقنية المعلومات، استُخدمت ٣٣٤٩ وثيقة، أي بنسبة ٨٤,٥٥٪، بوصفها corpus نهائيًا لنموذج LDA، في حين استُبعدت ٦١٢ وثيقة لعدم استيفائها معايير الاختبار. أُجريت عملية النمذجة باستخدام LDA مع اعتماد ٩ موضوعاتٍ نهائية، وحقق النموذج قيمة *coherence score* بلغت ٠,٤٢٩٧٢٣. وأظهرت نتائج التصنيف أنّ ٢٠٩٠ وثيقة قد صنّفت ضمن ستة مجالات في هندسة المعلوماتية، حيث كان مجال هندسة البرمجيات الأكثر هيمنةً بعدد ٥٨٣ وثيقة بنسبة ٢٧,٨٩٪، يليه مجال الأنظمة الذكية بعدد ٥٧٩ وثيقة بنسبة ٢٧,٧٠٪، ثمّ إنترنت الأشياء والروبوتات بعدد ٣٥٩ وثيقة بنسبة ١٧,١٨٪، فالوسائط المتعددة والألعاب بعدد ٢٨١ وثيقة بنسبة ١٣,٤٤٪، ثمّ نظم المعلومات بعدد ١٨٣ وثيقة بنسبة ٨,٧٦٪، وأخيرًا برمجة الويب والهواتف المحمولة بعدد ١٠٥ وثائق بنسبة ٥,٠٢٪. وتُظهر نتائج الدراسة أنّ نموذج LDA قادرٌ على تقديم تصوّر منظم لاتجاهات موضوعات رسائل التخرج لدى الطلاب، ويمكن الاستفادة منه مرجعًا في اختيار موضوعاتٍ أكثر ملاءمةً وحداثةً.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Skripsi merupakan karya ilmiah yang harus diselesaikan oleh mahasiswa sebagai syarat kelulusan pada tingkat akhir pendidikan. Selain menjadi media bagi mahasiswa untuk mengaplikasikan ilmu yang telah dipelajari selama masa perkuliahan, skripsi juga memiliki peran penting sebagai kontribusi akademik terhadap pengembangan ilmu pengetahuan di bidang studi masing-masing. Oleh karena itu, pemilihan topik skripsi yang tepat sangat menentukan kualitas penelitian yang dihasilkan. Topik yang relevan dan inovatif akan membuka peluang baru dalam pengembangan ilmu pengetahuan. Hal ini sejalan dengan penelitian oleh Creswell (2014) yang menyatakan bahwa pemilihan topik penelitian merupakan tahap krusial dalam menentukan arah dan kualitas penelitian. Selain itu, Booth et al. (2016) juga menegaskan bahwa topik yang baik harus memiliki kontribusi ilmiah dan relevansi terhadap perkembangan keilmuan.

Salah satu masalah utama yang sering dihadapi oleh mahasiswa dalam penyusunan skripsi adalah kesulitan dalam menentukan topik yang tepat. Berdasarkan berbagai survei akademik, mahasiswa sering kali merasa bingung dalam memilih topik yang sesuai dengan minat serta bidang studi mereka, namun tetap memiliki kontribusi ilmiah yang signifikan. Tidak jarang, topik yang dipilih dalam skripsi cenderung berulang dari tahun ke tahun dan kurang inovatif, yang mengurangi kualitas penelitian dan menghambat pengembangan ilmu pengetahuan. Sebagaimana yang ditunjukkan oleh Santoso et al. (2020), banyak mahasiswa yang

lebih cenderung memilih topik yang sudah umum, karena kurangnya eksplorasi terhadap topik baru. Temuan ini juga diperkuat oleh Hart (2018) yang menyatakan bahwa keterbatasan dalam eksplorasi literatur menyebabkan penelitian menjadi repetitif dan kurang inovatif.

Selain itu, kurangnya akses ke sumber informasi yang relevan dan sistem yang dapat membantu mahasiswa dalam mengeksplorasi topik-topik terkini juga menjadi kendala utama. Tanpa adanya platform atau teknologi yang mendukung, mahasiswa sering kali hanya mengandalkan saran dari dosen pembimbing atau meniru topik dari skripsi sebelumnya, yang menyebabkan topik penelitian menjadi terbatas dan tidak berkembang. Menurut Rowley (2012), keterbatasan akses terhadap sumber informasi ilmiah dapat menghambat proses pemilihan topik penelitian yang berkualitas. Selain itu, Nicholas et al. (2017) menjelaskan bahwa perilaku pencarian informasi mahasiswa cenderung terbatas pada sumber yang mudah diakses, sehingga mengurangi keberagaman ide penelitian.

Untuk mengatasi masalah tersebut, salah satu pendekatan yang digunakan adalah dengan menerapkan *Latent Dirichlet Allocation* (LDA), yaitu metode algoritma dalam *topic mining*. LDA adalah teknik statistik yang digunakan untuk mengidentifikasi dan mengungkap topik-topik tersembunyi dalam kumpulan dokumen besar, seperti arsip skripsi atau publikasi ilmiah. Dengan menggunakan LDA, data teks dapat dianalisis untuk menemukan pola atau tema yang sering muncul dalam dokumen tersebut, yang dapat memberikan wawasan bagi mahasiswa untuk memilih topik skripsi yang relevan, inovatif, dan belum banyak dieksplorasi. Hal ini didukung oleh penelitian David M. Blei et al. (2003) yang memperkenalkan LDA sebagai model probabilistik generatif untuk menemukan

struktur topik dalam dokumen. Selain itu, Steyvers & Griffiths (2007) juga menunjukkan bahwa LDA efektif dalam analisis teks skala besar.

Metode LDA memungkinkan peneliti untuk mengidentifikasi sejumlah topik berdasarkan distribusi kata-kata yang muncul dalam teks. Melalui pendekatan ini, mahasiswa tidak hanya dapat menemukan topik yang sesuai dengan minat dan bidang studi mereka, tetapi juga memperoleh wawasan tentang tren penelitian yang sedang berkembang di dunia akademik. Seperti yang dijelaskan dalam penelitian oleh Blei et al. (2003), LDA efektif digunakan untuk mengidentifikasi struktur topik dalam kumpulan data besar. Selain itu, Jeyasekaran & Sebastian (2020) menunjukkan bahwa topic modeling mampu mengidentifikasi tren penelitian secara otomatis dari kumpulan dokumen ilmiah.

Penerapan metode LDA dalam pemilihan topik skripsi tidak hanya memberikan keuntungan bagi mahasiswa, tetapi juga bagi institusi pendidikan yang dapat memetakan tren-topik penelitian yang sedang berkembang. Dengan demikian, LDA menjadi alat yang sangat berguna dalam meningkatkan efisiensi pemilihan topik, yang pada akhirnya dapat menghasilkan skripsi dengan kualitas yang lebih baik dan relevansi yang lebih tinggi terhadap perkembangan ilmu pengetahuan. Penelitian oleh Alghamdi & Alfalqi (2015) menunjukkan bahwa topic modeling, khususnya LDA, mampu memberikan insight yang signifikan dalam analisis dokumen teks. Selain itu, Syed & Spruit (2017) menyatakan bahwa LDA banyak digunakan dalam berbagai domain untuk analisis tren dan pengambilan keputusan berbasis data.

Hal ini sejalan dengan firman Allah dalam QS. Al-‘Alaq ayat 1–5, yang menekankan pentingnya membaca, memahami, dan menuntut ilmu:

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ﴿١﴾ خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ ﴿٢﴾ اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ﴿٣﴾ الَّذِي عَلَّمَ بِالْقَلَمِ ﴿٤﴾ عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ

“Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan. Dia telah menciptakan manusia dari segumpal darah. Bacalah, dan Tuhanmulah Yang Maha Mulia, Yang mengajar (manusia) dengan pena, Dia mengajarkan manusia apa yang tidak diketahuinya.”(QS. Al-Alaq: 1-5).

Tafsir QS. Al-‘Alaq ayat 1–5 menurut Tafsir Ibnu Katsir menekankan pentingnya aktivitas membaca, proses pembelajaran, serta penyampaian ilmu melalui tulisan sebagai bentuk anugerah Allah kepada manusia (Mukmin, 2016). Berdasarkan penjelasan tersebut, ilmu pengetahuan dapat dipahami sebagai sesuatu yang perlu dicari, dipelajari, dan dikembangkan agar memberikan manfaat yang lebih luas. Dalam konteks penelitian skripsi, penerapan metode *Latent Dirichlet Allocation* (LDA) dapat dimaknai sebagai bentuk ikhtiar untuk “membaca” dan menggali informasi yang tersembunyi dalam kumpulan data teks, sehingga memungkinkan mahasiswa menemukan topik-topik baru yang relevan bagi pengembangan ilmu pengetahuan. Dengan demikian, pemanfaatan teknologi seperti LDA mencerminkan upaya manusia dalam memahami, mengolah, dan menghasilkan pengetahuan yang sebelumnya belum diketahui.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang tersebut, rumusan masalah dalam penelitian ini adalah bagaimana menghasilkan secara otomatis proses identifikasi dan pemetaan topik skripsi mahasiswa dengan menggunakan metode *Topic mining* berbasis *Latent Dirichlet Allocation*(LDA)?

1.3 Batasan Masalah

Penulis membatasi masalah pada penelitian ini sebagai berikut:

1. Penelitian ini hanya menggunakan data skripsi yang diambil dari *etheses* UIN Maulana Malik Ibrahim Malang dan dalam rentang waktu 2020-2025 dan berfokus pada Program Studi Teknik Informatika.
2. Analisis terbatas pada identifikasi topik berdasarkan judul, abstrak, dan kata kunci skripsi tanpa mengkaji lebih lanjut isi skripsi secara mendalam.
3. Metode yang digunakan adalah *Latent Dirichlet Allocation* (LDA), yang hasilnya bergantung pada efektivitas model dalam mengidentifikasi dan pemetaan tren topik mahasiswa.
4. Penelitian ini tidak mencakup pengembangan sistem informasi berbasis *topic mining* yang terintegrasi.

1.4 Tujuan Penelitian

Berdasarkan pernyataan masalah, tujuan penelitian ini adalah untuk mengidentifikasi tren penelitian dan memetakan evolusi topik penelitian yang telah dibahas oleh mahasiswa dalam karya ilmiah berupa skripsi. Selain itu Penelitian ini juga diharapkan dapat memberikan solusi atas tantangan yang sering dihadapi mahasiswa dalam menentukan topik skripsi yang relevan, inovatif, dan memiliki nilai akademik tinggi.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah dapat memperbaiki kualitas perangkat lunak dengan mendeteksi cacat lebih awal, sehingga mengurangi biaya dan waktu perbaikan, menjadi referensi bagi penelitian

lanjutan di bidang prediksi cacat perangkat lunak dan machine learning, dan membantu perusahaan meningkatkan keandalan perangkat lunak dan mengoptimalkan sumber daya.

BAB II

STUDI PUSTAKA

Pembahasan dalam bab ini mencakup beberapa aspek penting, yaitu penelitian terdahulu, konsep skripsi mahasiswa, pengertian dan proses topic mining, metode Latent Dirichlet Allocation (LDA), serta tahapan text processing yang digunakan dalam penelitian ini. Dengan adanya kajian pustaka yang sistematis, diharapkan penelitian yang dilakukan memiliki dasar ilmiah yang kuat, serta mampu menunjukkan posisi dan kontribusi penelitian ini dibandingkan dengan penelitian sebelumnya.

2.1 Penelitian Terdahulu

Pada penelitian ini, pemilihan topik skripsi menjadi salah satu aspek penting dalam menentukan kualitas dan relevansi penelitian yang dilakukan oleh mahasiswa. Banyak mahasiswa menghadapi kesulitan dalam menemukan topik skripsi yang inovatif dan sesuai dengan perkembangan ilmu pengetahuan, serta relevansi praktis di bidang studi mereka. Salah satu solusi yang ditawarkan untuk mengatasi masalah ini adalah dengan memanfaatkan metode *topic mining*, yang memungkinkan identifikasi topik-topik potensial berdasarkan analisis data skripsi sebelumnya. Bab ini akan membahas berbagai penelitian terdahulu yang berkaitan dengan penerapan *topic mining*, serta kontribusinya dalam membantu mahasiswa dalam mengeksplorasi topik yang lebih relevan dan berinovasi, sekaligus memberikan dampak positif terhadap pengembangan ilmu pengetahuan.

Pada penelitian oleh (Asmussen & Møller, 2019), *Latent Dirichlet Allocation* (LDA) diterapkan dalam sebuah *framework* untuk melakukan tinjauan literatur eksploratif pada bidang analitik, manajemen rantai pasokan (SCM), dan sistem informasi perusahaan. Dalam studi ini, jumlah dokumen yang ditemukan mencapai 650, namun dengan menggunakan *framework* berbasis LDA, jumlah tersebut berhasil dipersempit menjadi hanya 76 dokumen yang relevan, yang menegaskan kemampuan LDA dalam mengelompokkan dokumen secara efisien. Keberhasilan ini menunjukkan bahwa LDA, sebagai metode *unsupervised learning*, sangat efektif dalam menangani kumpulan dokumen besar tanpa membutuhkan pengetahuan atau kategori sebelumnya.

Proses ini menghemat waktu yang diperlukan untuk membaca dan menganalisis setiap dokumen secara manual, yang biasanya memakan banyak waktu dan sumber daya, terutama dalam penelitian dengan jumlah dokumen yang sangat besar. Dengan mengurangi jumlah dokumen yang harus ditinjau secara manual, LDA memungkinkan peneliti untuk fokus pada topik-topik utama dan penelitian yang relevan, tanpa harus terjebak dalam jumlah dokumen yang terlalu banyak.

Dalam konteks *topic mining* untuk skripsi mahasiswa, kemampuan LDA untuk mengidentifikasi dan mengelompokkan topik-topik secara otomatis dapat meningkatkan akurasi tinjauan literatur serta memberikan gambaran yang lebih jelas tentang arah penelitian yang sedang berkembang. Penggunaan LDA dalam penelitian ini juga membuktikan bahwa *framework* ini tidak hanya efisien tetapi juga dapat diulang dan transparan, yang memungkinkan hasil tinjauan literatur yang

konsisten dan lebih cepat dibandingkan metode manual atau pendekatan lain yang lebih memakan waktu.

Dalam penelitian (Zhang *et al.*, 2020), metode *multi-dimensional topic mining* yang berbasis pada *hierarchical semantic graph* digunakan untuk mengidentifikasi topik-topik dalam berbagai dokumen. Model ini menggabungkan dua pendekatan utama, yaitu *co-occurrence* dan *word2vec*, untuk membangun sebuah grafik semantik yang mengungkap hubungan antar istilah dengan mempertimbangkan konteks dan konten dokumen. Keunggulan utama dari pendekatan ini adalah kemampuannya untuk menggali struktur hierarkis dari istilah dalam dokumen, menghasilkan representasi yang lebih akurat dan kaya dari topik-topik yang ada.

Dengan menerapkan teknik analisis struktur grafis dan segmentasi subgraf, model ini dapat mendeteksi fitur multi-dimensional dari topik, yang membantu dalam mengungkap keterkaitan antara istilah dan topik secara lebih jelas. Penelitian ini menunjukkan bahwa dengan menggunakan model grafik semantik yang diperbarui, yang menggabungkan *content correlation* dan *contextual correlation*, mereka berhasil mendapatkan representasi yang lebih komprehensif tentang hubungan antara istilah dalam dokumen. Hasil eksperimen yang dilakukan pada dataset dokumen besar menunjukkan bahwa metode ini mampu mengungguli metode-metode terkini dalam hal akurasi dan efisiensi. Dalam eksperimen ini, hasil *topic mining* menunjukkan peningkatan yang signifikan dalam kualitas topik yang diekstraksi, dengan tingkat akurasi mencapai 80-85% dalam beberapa percobaan pada dataset besar, yang mencakup ribuan dokumen.

Selanjutnya, Dalam penelitian (Cai et al., 2019), sebuah metode *topic mining semi-supervised* yang berbasis pada *hierarchical clustering* digunakan untuk menangani *multi-source network public opinion*. Metode ini bertujuan untuk menggali topik dan struktur hierarkis dari teks pendek yang banyak terdapat di platform jejaring sosial dan komunitas online. Salah satu inovasi penting dari penelitian ini adalah peningkatan pada algoritma *clustering* hierarkis tradisional yang mengoptimalkan input vektorisasi model dan pemilihan topik berkualitas tinggi, serta kemampuan untuk menangani teks dengan banyak kebisingan dan norma gramatikal yang tidak terdefinisi dengan baik. Dalam eksperimen yang dilakukan, data yang digunakan berasal dari forum *Tsinghua* dan BUPT (*Beijing University of Posts and Telecommunications*), dengan pengumpulan data menggunakan teknologi *web crawler*.

Penelitian ini menunjukkan bahwa model yang ditingkatkan ini dapat mengekstrak topik dan hubungan hierarkis antar topik dengan baik. Hasil percakapan pertama menunjukkan bahwa model yang ditingkatkan memiliki tingkat akurasi dan *recall rate* yang lebih tinggi, terutama dalam pengelompokan topik yang lebih halus, yang menunjukkan efektivitasnya dalam memproses data teks pendek yang banyak terdapat kebisingan. Model ini juga dapat menggali lebih dalam mengenai sub-topik dan hirarki topik dalam dataset yang besar, memberikan keuntungan dalam analisis lanjutan seperti retrieval tema dan evolusi tema. Hasil eksperimen menunjukkan bahwa metode ini mengungguli dalam hal penambahan topik yang lebih terperinci, dengan akurasi mencapai 73% dan *recall rate* yang lebih tinggi, khususnya dalam pengklasifikasian tingkat kedua dari topik. Ini

menjadikan metode ini sangat relevan untuk analisis teks pendek yang banyak ditemukan di platform multi-sumber, yang memiliki berbagai topik dan banyak teks yang mengandung kebisingan. Pendekatan ini memiliki potensi besar untuk digunakan dalam pemodelan opini publik di platform jejaring sosial.

Dalam penelitian (Shah et al., 2021), pendekatan *topic mining* menggunakan model *Latent Dirichlet Allocation* (LDA) yang ditingkatkan, yaitu TCLDA (*Topic Coherence -based LDA*), digunakan untuk menganalisis topik dan dinamika sentimen pada ulasan pasien mengenai dokter di platform rating dokter online selama gelombang awal pandemi *COVID-19*. Data yang digunakan dalam penelitian ini berasal dari 55.612 ulasan pasien mengenai 3430 dokter dari tiga platform terkemuka di AS (*RateMDs*, *HealthGrades*, dan *Vitals*), yang dikumpulkan antara 1 Maret hingga 27 Juni 2020. Melalui penerapan TCLDA, mereka berhasil mengidentifikasi 30 topik yang terkait dengan perawatan medis selama pandemi, yang kemudian dikelompokkan berdasarkan kategori penyakit dengan tingkat keparahan yang berbeda (penyakit dengan tingkat kematian tinggi dan rendah). Selain itu, metode ini memungkinkan pengidentifikasian topik yang sedang muncul dan memudar seiring waktu, yang penting untuk memahami dinamika persepsi pasien mengenai kualitas pelayanan kesehatan di masa krisis. Hasil eksperimen menunjukkan bahwa model TCLDA mampu memberikan akurasi yang lebih tinggi dalam hal identifikasi topik yang lebih halus dan hubungan hierarkis antar topik.

Penelitian ini menemukan bahwa emosi negatif, seperti ketakutan, kemarahan, dan kesedihan, mendominasi ulasan pasien selama gelombang awal

pandemi *COVID-19*, mencerminkan ketidakpastian dan ketakutan terkait dengan pandemi. Selain itu, penelitian ini juga menyoroti pentingnya sentimen dalam ulasan pasien sebagai elemen penting dalam memahami pengalaman dan kebutuhan mereka, yang pada gilirannya dapat membantu meningkatkan kualitas pelayanan kesehatan. TCLDA terbukti efektif dalam menangani teks pendek yang kaya akan kebisingan dan ketidakpastian, seperti yang terdapat dalam ulasan pasien, memberikan wawasan yang lebih mendalam tentang topik-topik terkait dengan pengalaman pasien dan emosi mereka selama pandemi.

Penelitian lain oleh Rashid et al. (2019) mengusulkan sebuah teknik *topic modeling* yang menggabungkan *inverse document frequency* dan *fuzzy k-means clustering* untuk penambahan teks pada dokumen biomedis. Metode ini dirancang untuk mengatasi masalah redundansi dan dimensi tinggi yang sering muncul dalam pengolahan teks biomedis. Dalam pendekatan ini, *hybrid inverse document frequency* (IDF) digunakan untuk menilai bobot istilah secara global, sedangkan *fuzzy k-means clustering* diterapkan untuk mengelompokkan dokumen dengan cara yang lebih fleksibel, di mana sebuah data dapat memiliki keanggotaan parsial pada beberapa *cluster* atau topik. Teknik ini terbukti membantu mengurangi redundansi pada dokumen biomedis yang umumnya mengandung informasi berulang, serta menangani kompleksitas akibat dimensi tinggi dalam data teks medis.

Hasil penelitian menunjukkan bahwa pendekatan yang diusulkan menghasilkan akurasi yang lebih tinggi dalam tugas klasifikasi dan klusterisasi dibandingkan dengan model topik lain seperti LSA. Dalam eksperimen menggunakan dataset biomedis, akurasi klasifikasi mencapai 92,35% pada dataset

Ohsumed dan 98,97% pada dataset *MuchMore Springer Corpus*. Selain itu, metode ini menunjukkan waktu eksekusi yang stabil meskipun jumlah topik yang dianalisis meningkat, menandakan efisiensi yang unggul dibandingkan metode lain yang lebih kompleks. Dengan mengombinasikan teknik reduksi dimensi menggunakan *principal component analysis* (PCA), metode ini berhasil meningkatkan kualitas pengelompokan dan klasifikasi, serta mengatasi tantangan umum dalam penambangan teks biomedis, seperti redundansi dan dimensi tinggi.

Pendekatan inovatif lain dalam penelitian yang dilakukan (Sun et al., 2020), *Sentimental Image Dominant Graph Topic Model* (SIDGTM) digunakan untuk mendeteksi topik dari data heterogen yang melibatkan berbagai jenis media seperti teks dan gambar dalam konteks jaringan 5G. Model ini menggabungkan analisis sentimen pada data visual dan teks untuk memahami topik secara lebih mendalam, dengan mengoptimalkan pemodelan topik menggunakan jaringan grafis berbobot yang menghubungkan gambar dan teks. SIDGTM bertujuan untuk menjembatani kesenjangan semantik yang ada antara data teks dan gambar, sehingga memungkinkan deteksi topik yang lebih akurat dari data yang memiliki berbagai modalitas, seperti yang ditemukan dalam aplikasi berbasis 5G yang sangat bergantung pada data visual dan teks. Melalui penggunaan *Convolutional Neural Networks* (CNN), SIDGTM meningkatkan kekuatan diskriminasi fitur visual, yang memungkinkan model ini untuk mengekstrak dan menghubungkan informasi visual yang relevan dengan topik yang ada dalam teks. Hasil eksperimen menunjukkan bahwa SIDGTM secara signifikan mengungguli model-model lain dalam hal akurasi deteksi topik pada *dataset cross-modality*, seperti yang diujikan pada

dataset MCG-WEBV. Model ini menunjukkan *F-measure* yang lebih tinggi, menunjukkan bahwa SIDGTM lebih efisien dalam mengekstrak topik dari data yang menggabungkan teks dan gambar. Selain itu, penelitian ini juga mengimplementasikan analisis sentimen untuk menilai polaritas sentimen dalam topik yang terdeteksi, menggunakan skor semantik berdasarkan *SentiWordNet*, yang memungkinkan model untuk mendeteksi sentimen positif, negatif, atau netral dari setiap topik. Total akurasi deteksi topik yang dicapai oleh SIDGTM dalam eksperimen ini adalah 87%.

Tabel 2. 1 Penelitian terkait

No	Referensi	Penulis	Metode Penelitian	Hasil Penelitian
1.	<i>Smart literature review: a practical topic modelling approach to exploratory literature review</i>	Asmussen & Møller	Pada Penelitian ini menggunakan metode <i>Latent Dirichlet Allocation</i> (LDA), sebuah metode topic modeling, untuk melakukan tinjauan literatur eksploratif.	kerangka kerja tinjauan literatur eksploratif yang menggunakan <i>Latent Dirichlet Allocation</i> (LDA) untuk mengelompokkan 650 paper, kerangka kerja ini berhasil mengelompokkan paper-paper tersebut ke dalam 20 topik yang relevan.
2.	<i>Multi-Dimension Topic Mining Based on Hierarchical Semantic Graph Model</i>	Zhang et al.	Pada penelitian ini menggunakan metode <i>hierarchical semantic graph</i> dengan menggabungkan dua pendekatan utama, yaitu <i>co-occurrence</i> dan <i>word2vec</i> .	Hasil akurasi mencapai 80-85% dalam beberapa percobaan pada dataset besar, yang mencakup ribuan dokumen.
3.	<i>A Topic mining Method for Multi-source Network Public Opinion Based on Improved Hierarchical Clustering</i>	Cai et al.	Pada penelitian ini menggunakan metode <i>topic mining</i> <i>semi-supervised</i> yang berbasis pada <i>hierarchical clustering</i> .	Hasil akurasi mencapai 73% dan recall rate yang lebih tinggi, khususnya dalam pengklasifikasian tingkat kedua dari topik.

4.	<i>Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach</i>	Shah et al.	Pada penelitian ini menerapkan metode TCDL A (<i>Topic Coherence - based Latent Dirichlet Allocation</i>)	Hasil dari penelitian ini menemukan bahwa emosi negatif, seperti ketakutan, kemarahan, dan kesedihan, mendominasi ulasan pasien selama gelombang awal pandemi COVID-19, mencerminkan ketidakpastian dan ketakutan terkait dengan pandemi.
5.	<i>Topic Modeling Technique for Text Mining Over Biomedical Text Corpora through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering</i>	Rashid et al.	Pada penelitian ini menerapkan <i>topic modeling</i> yang menggabungkan <i>inverse document frequency</i> dan <i>fuzzy k-means clustering</i>	hasil klasifikasi menunjukkan peningkatan akurasi yang signifikan, dengan akurasi mencapai hingga 92,35% pada dataset <i>Ohsumed</i> dan 98,97% pada dataset <i>MuchMore Springer Corpus</i> .
6.	<i>Enabling 5G: sentimental image dominant graph topic model for cross-modality topic detection</i>	Sun et al.	Pada penelitian ini menggunakan metode <i>Sentimental Image Dominant Graph Topic Model (SIDGTM)</i>	Hasil dari metode yang diusulkan mencapai akurasi deteksi topik yang dicapai oleh SIDGTM dalam eksperimen ini adalah 87%.
7.	<i>Systematic Literature Review untuk Mengidentifikasi dan Mengklasifikasi Metode Topic mining</i>	Ahmad Arsha Albar, M. Ainul Yaqin	Pada penelitian ini menggunakan <i>Systematic literature Review (SLR)</i> dengan kerangka kerja PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).	Hasil dari penelitian ini yaitu untuk mengetahui tren publikasi yang terus meningkat, rekomendasi untuk penelitian masa depan seperti pengembangan LDA.

Berdasarkan Tabel 2.1, penelitian-penelitian terdahulu menunjukkan bahwa metode *topic mining* telah banyak diterapkan dalam berbagai domain dengan pendekatan yang beragam. Penelitian oleh Asmussen & Møller (2019) memanfaatkan *Latent Dirichlet Allocation* (LDA) untuk melakukan tinjauan literatur eksploratif, sedangkan Zhang et al. (2020) menggunakan pendekatan hierarchical semantic graph untuk meningkatkan representasi hubungan antar topik.

Di sisi lain, Cai et al. (2019) menerapkan hierarchical clustering untuk menangani teks pendek, sementara Shah et al. (2021) mengembangkan TCLDA untuk meningkatkan kualitas topik dan analisis sentimen. Selain itu, Rashid et al. (2019) menggunakan kombinasi IDF dan fuzzy k-means clustering untuk klasifikasi dokumen biomedis, serta Sun et al. (2020) mengembangkan model multimodal yang menggabungkan teks dan gambar dalam proses *topic mining*. Adapun penelitian oleh Ahmad Arsha Albar & M. Ainul Yaqin (2025) menggunakan pendekatan Systematic Literature Review (SLR) untuk mengidentifikasi dan mengklasifikasikan metode *topic mining* tanpa melakukan implementasi langsung pada dataset tertentu.

Meskipun penelitian-penelitian tersebut menunjukkan kemajuan yang signifikan, terdapat perbedaan mendasar dengan penelitian yang akan dilakukan. Pertama, dari segi objek penelitian, sebagian besar penelitian terdahulu menggunakan data umum seperti artikel ilmiah, ulasan daring, media sosial, dan dokumen biomedis, sedangkan penelitian ini secara khusus menggunakan data skripsi mahasiswa UIN Maulana Malik Ibrahim Malang sebagai objek utama. Hal ini menjadikan penelitian ini lebih kontekstual dalam lingkungan akademik serta memiliki tujuan yang lebih spesifik, yaitu untuk memetakan topik skripsi mahasiswa.

Dengan demikian, dapat disimpulkan bahwa penelitian ini memiliki perbedaan yang signifikan dibandingkan penelitian terdahulu, baik dari segi objek, metode, maupun tujuan penelitian. Penelitian ini tidak hanya mengadopsi metode yang telah ada, tetapi juga mengarah pada penerapan yang lebih spesifik dan

aplikatif, sehingga diharapkan dapat memberikan kontribusi baru dalam bidang *topic mining*, khususnya dalam pemetaan tren topik skripsi mahasiswa.

2.2 *Topic mining*

Topic mining adalah teknik dalam *data mining* yang digunakan untuk mengidentifikasi topik utama atau tema tersembunyi dalam kumpulan data teks besar, seperti dokumen, artikel, atau skripsi mahasiswa. Teknik ini memanfaatkan algoritma untuk menganalisis data teks secara otomatis guna menemukan pola-pola yang mengungkapkan topik yang relevan serta tren atau perkembangan terbaru dalam suatu bidang keilmuan. *Topic mining*, yang awalnya hanya mengandalkan statistik frekuensi kata, kini telah berkembang pesat menjadi pendekatan berbasis *deep learning* yang lebih canggih dalam mengekstraksi tema dan pola evaluasi dari data teks (Meng et al., 2024). Dalam konteks pemilihan topik skripsi, *topic mining* dapat membantu mahasiswa dan dosen pembimbing menemukan topik yang inovatif dan sesuai berdasarkan analisis data yang objektif (Nugraha et al., 2024).

Proses *topic mining* melibatkan beberapa langkah, yang dimulai dengan tahapan *preprocessing* data teks. Pada tahap ini, beberapa langkah penting dilakukan untuk mempersiapkan data yang akan dianalisis, yang meliputi:

- a) Penghilangan *Stopwords*: *Stopwords* adalah kata-kata yang sering muncul dalam teks tetapi tidak memberikan informasi penting untuk analisis, seperti "dan," "atau," "di," dll. Menghapus kata-kata ini membantu dalam fokus pada kata-kata kunci yang lebih bermakna.
- b) Tokenisasi: Tokenisasi adalah proses memecah teks menjadi unit yang lebih kecil, biasanya kata atau frasa, yang disebut token. Tokenisasi

memungkinkan pemisahan kata-kata dalam kalimat untuk mempermudah analisis lebih lanjut.

- c) *Stemming* dan *Lemmatization*: *Stemming* adalah proses mengubah kata menjadi bentuk dasar atau akarnya, sedangkan *lemmatization* adalah proses untuk mengembalikan kata ke bentuk dasar secara linguistik, misalnya mengubah "berlari" menjadi "lari". Kedua teknik ini membantu dalam mengurangi variasi kata yang memiliki makna sama.
- d) Penghapusan Tanda Baca dan Angka: Dalam tahap ini, tanda baca dan angka yang tidak relevan dengan analisis dikeluarkan dari teks untuk mengurangi gangguan dalam pemrosesan selanjutnya.
- e) *Lowercasing*: Semua teks biasanya diubah menjadi huruf kecil untuk memastikan bahwa analisis tidak dibedakan berdasarkan kapitalisasi (misalnya, "Data" dan "data" dianggap sama).

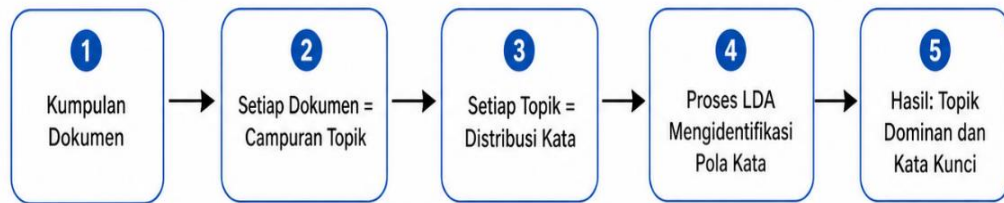
Setelah tahap *preprocessing*, langkah selanjutnya adalah identifikasi kata kunci yang sering muncul, serta pengelompokan dokumen berdasarkan kesamaan tema. Dalam konteks penelitian skripsi, *topic mining* memiliki berbagai manfaat. Salah satunya adalah membantu mahasiswa menemukan topik penelitian yang terarah, relevan, dan belum banyak dibahas sebelumnya. Hal ini dapat mengurangi kemungkinan pemilihan topik yang sudah umum atau terulang, yang seringkali mengarah pada penelitian yang kurang inovatif. Selain itu, dengan menggunakan *topic mining*, mahasiswa dapat melihat tren dan perkembangan terbaru dalam bidang studi tertentu, yang mungkin tidak mudah ditemukan hanya dengan pencarian manual atau berdasarkan skripsi-skripsi sebelumnya.

Bagi institusi pendidikan, penerapan *topic mining* juga memberikan keuntungan. Dengan memetakan topik-topik yang sedang tren di kalangan mahasiswa, perguruan tinggi dapat lebih mudah mengidentifikasi area yang perlu dieksplorasi lebih lanjut dalam penelitian. Selain itu, *topic mining* dapat membantu dalam mengelola dan mengarsipkan skripsi mahasiswa dengan cara yang lebih terstruktur, sehingga memudahkan akses dan pemantauan perkembangan penelitian dari waktu ke waktu.

Dengan demikian, *topic mining* tidak hanya memudahkan mahasiswa dalam menentukan topik skripsi, tetapi juga memberikan kontribusi signifikan terhadap perkembangan ilmu pengetahuan di tingkat institusi pendidikan. Implementasi metode ini membuka peluang baru untuk meningkatkan efisiensi dan kualitas penelitian akademik, serta menciptakan inovasi dalam penelitian yang lebih relevan dengan perkembangan terkini di bidangnya.

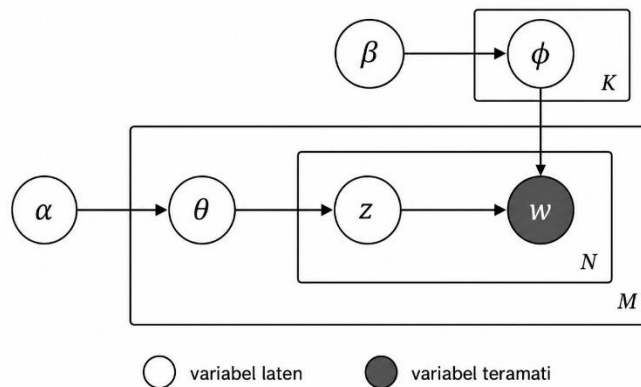
2.3 *Latent Dirichlet Allocation (LDA)*

LDA (*Latent Dirichlet Allocation*) adalah metode analisis topik yang dapat digunakan untuk menemukan topik tersembunyi dalam kumpulan dokumen (Setijohatmo et al., 2020). LDA mengasumsikan bahwa setiap dokumen terdiri dari campuran beberapa topik secara proporsional, dan setiap topik terdiri dari campuran beberapa kata secara proporsional. Ilustrasi pembentukan kata pada topik Gambar 2.1.

Gambar 2. 1 Ilustrasi *Latent Dirichlet Allocation*

Dalam implementasinya, LDA tidak dapat diterapkan karena variabel z , θ dan ϕ tersembunyi/tidak diketahui. Pada *plate* dapat dilihat untuk gambar yang diarsir adalah variabel yang teramati yaitu w , dan variabel lainnya yang tidak diarsir adalah variabel tersembunyi. Dalam proses LDA diatas diasumsikan dengan perulangan generative yang dilakukan untuk setiap dokumen M dalam sebuah *Corpus*.

Seperti yang ditunjukkan pada gambar berikut, LDA dijelaskan oleh model grafis menggunakan notasi plat pada gambar 2.2.



Gambar 2. 2 Plat Notasi LDA

Keterangan:

- α adalah parameter Dirichlet distribusi dari topik terhadap dokumen
- β adalah parameter Dirichlet distribusi dari kata terhadap topik
- ϕ_k adalah distribusi kata untuk topik k
- θ_m adalah distribusi topik untuk dokumen m
- z_{mn} adalah topik untuk kata ke- n pada dokumen m
- w_{mn} adalah kata pada urutan tertentu ke- n pada dokumen m

N adalah himpunan kata yang berada dalam dokumen (dokumen m memiliki N_m kata),
dan
 M adalah Kumpulan dari dokumen.

Berdasarkan model, probabilitas dari model pada dokumen adalah sebagai berikut:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{m=1}^M P(\alpha) \prod_{k=1}^K P(\beta) \prod_{n=1}^{N_m} P(\theta_m) P(\phi_{z_{m,n}}) \quad (2.1)$$

Keterangan:

w : kata dalam dokumen
 Z : topik dari setiap kata
 θ : distribusi topik pada dokumen
 ϕ : distribusi kata pada topik
 α : parameter Dirichlet untuk distribusi topik-dokumen
 β : parameter Dirichlet untuk distribusi kata-topik
 M : jumlah dokumen dalam *corpus*
 K : jumlah topik
 N_m : jumlah kata pada dokumen ke- m

Pada persamaan (2.1) diasumsikan bahwa topik setiap dokumen mengikuti distribusi *Dirichlet* pada persamaan (2.2):

$$P(\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \Gamma(\theta_k^{\alpha_k - 1}) \quad (2.2)$$

Keterangan:

$P(\theta)$: distribusi probabilitas topik pada dokumen
 α : parameter Dirichlet (hyperparameter)
 $\Gamma(\alpha)$: fungsi gamma (generalized factorial)
 θ_k : proporsi topik ke- k dalam dokumen

Teruntuk distribusi peluang dari z untuk semua dokumen dan topik dalam *terms*, ditunjukkan pada persamaan (2.3) oleh n_m , dan k adalah jumlah topik k yang termasuk dalam kelompok kata dalam dokumen m :

$$P(\theta) = m = 1Mk = 1K\theta_{m,k}^{n_{m,k}} \quad (2.3)$$

Keterangan:

θ_m : distribusi topik pada dokumen ke- m
 $n_{m,k}$: jumlah kata dalam dokumen m yang termasuk topik k
 K : jumlah topik

Distribusi peluang bersyarat untuk semua *Corpus* ϕ_k juga mengikuti distribusi *Dirichlet* dengan parameter $\phi_{k,v}$ yang diformulasikan pada persamaan (2.4):

$$P(\beta) = \prod_{k=1}^K \frac{\Gamma(\beta_{kv})}{\prod_{v=1}^V \Gamma(\beta_{kv})} \prod_{v=1}^V \phi_{kv}^{\beta_{kv}-1} \quad (2.4)$$

Keterangan:

- ϕ_k : distribusi kata pada topik ke- k
- β : parameter *Dirichlet* untuk kata-topik
- v : indeks kata dalam vocabulary
- V : jumlah seluruh kata unik (*vocabulary size*)

Dalam menentukan pembagian kata atau distribusi kata dalam pencarian Z_{mn} , LDA menggunakan *gibbs sampling*. *Gibbs sampling* adalah salah satu algoritma keluarga dari *Markov Chain Monte Carlo* (MCMC). *Gibbs sampling* akan *sample* satu per satu kata pada setiap dokumen untuk pemberian topik untuk menghubungkan variabel satu dengan variabel lain. Dalam proses *gibbs sampling* akan dilakukan pengelompokan topik dimana proses yang perlu dilakukan yaitu:

1. Distribusi probabilitas topik pada suatu dokumen

Pada proses probabilitas topik pada suatu dokumen, dimisalkan terdapat n topik yang muncul pada suatu dokumen d , yang ditunjukkan pada persamaan (2.5):

$$n_{d,k} + \alpha \quad (2.5)$$

Keterangan:

- $n_{d,k}$: jumlah kata dalam dokumen d yang masuk topik k
- α : hyperparameter *Dirichlet*
- z : topik

2. Distribusi probabilitas kata pada suatu topik

Dalam pencarian distribusi probabilitas kata pada topik adalah dengan mengetahui berapa banyak m kata w muncul pada topik k pada *Corpus*. menghasilkan persamaan (2.6):

$$\frac{m_{w,k} + \beta}{\sum_{i=1}^V \sum m_k + V\beta} \quad (2.6)$$

Keterangan:

$m_{w,k}$: jumlah kemunculan kata w pada topik k
 $\sum m_k$: total kata dalam topik k
 V : jumlah *vocabulary*
 β : *hyperparameter*
 w : kata

Mirip dengan persamaan sebelumnya, dimana $m_{w,k}$ adalah banyaknya kemunculan kata w yang muncul pada topik k pada seluruh korpus. W adalah banyak *term* kata yang ada pada dokumen. β adalah *hyperparameter*. Dan m_k adalah banyaknya kata yang tergabung pada topik k pada seluruh *Corpus*.

2.4 Text Processing

Text preprocessing merupakan tahap awal yang sangat penting dalam proses *natural language processing* (NLP). Tahap ini berfungsi untuk membersihkan data masukan yang masih tidak terstruktur dengan cara menghilangkan elemen-elemen yang dianggap tidak relevan. Beberapa teknik utama yang umum digunakan dalam *text preprocessing* antara lain *tokenization*, *Stopword Removal*, serta *Stemming* (Chai, 2023).

Text Processing adalah tahap yang perlu dilakukan sebelum membuat model pelatihan. Proses ini membersihkan data input yang masih kotor, Dimana banyak kata yang tidak perlu digunakan akan dibersihkan. Proses yang akan digunakan pada *Text Processing* yaitu *Tokenizing*, *Stopword*, *Stemming*.

Tokenizing adalah proses membagi teks menjadi unit yang lebih kecil, yang berupa kata atau frasa. *Tokenizing* adalah langkah penting dalam *text mining* karena memungkinkan untuk memisahkan setiap kata ke dalam unit-unit kecil dalam suatu array atau *term* (Togatorop et al., 2021). Pada *Tokenizing* juga diawali dengan *lowercasing* dimana teks yang ada akan diubah menjadi huruf kecil untuk memberikan hasil lebih maksimal.

Stopword adalah kata yang muncul dalam secara berulang pada teks tetapi tidak memberikan informasi penting atau spesifik. Seperti kata depan (di, pada, dari), kata hubung (dan, atau), kata ganti (saya, dia, mereka), dll. Dalam *text mining*, kata-kata berhenti biasanya dihapus dari teks sebelum analisis lebih lanjut. Tujuannya adalah untuk mengurangi ukuran data, mempercepat proses pengindeksan, dan meningkatkan relevansi hasil pencarian (Nugraha et al., 2017).

Stemming adalah proses menghasilkan perubahan morfologis dari akar kata. *Stemming* adalah teknik yang digunakan untuk menormalkannya agar lebih mudah untuk diproses.

2.5 Skripsi Mahasiswa

Tugas akhir atau yang biasa disebut dengan skripsi merupakan salah satu matakuliah yang wajib diambil oleh mahasiswa tingkat akhir. Skripsi merupakan sebuah karya ilmiah yang ditulis oleh mahasiswa tingkat akhir yang membahas tentang bidang atau topik tertentu berdasarkan dari suatu hasil kajian pustaka yang ditulis oleh para ahli, hasil pengembangan, atau hasil penelitian lapangan (Atmaja & Mandala, 2020).

Penulisan skripsi merupakan proses kompleks yang melibatkan berbagai tahapan pembelajaran, mulai dari pemahaman metodologi, pengumpulan data, analisis, hingga pembahasan secara sistematis dan terstruktur (Ismanto et al., 2023). Dalam proses tersebut, kajian pustaka tidak sekadar menjadi daftar referensi, tetapi harus memberikan ulasan kritis terhadap literatur yang relevan untuk memperdalam penelitian (Hadi & Afandi, 2021). Selain itu, penulisan skripsi menuntut kemampuan literasi informasi, pengelolaan data, serta kolaborasi dengan pihak lain sebagai bagian penting dari proses pembelajaran (Septiono, 2021). Lebih jauh, aspek berpikir kritis seperti interpretasi, analisis, dan inferensi diperlukan agar mahasiswa dapat menyumbangkan gagasan dan menawarkan solusi terhadap permasalahan yang diteliti (Nurcaya, 2021).

Dengan kompleksitas tersebut, skripsi tidak hanya menjadi latihan akademik, tetapi juga sarana bagi mahasiswa untuk mengidentifikasi masalah di masyarakat atau bidang ilmu tertentu dan memberikan solusi berdasarkan temuan mereka. Bentuk penelitian dalam skripsi dapat berupa penelitian dasar, penelitian terapan, maupun penelitian pengembangan yang berorientasi pada inovasi dan pemecahan masalah praktis. Referensi yang digunakan dalam skripsi harus valid, relevan, dan berasal dari sumber yang kredibel seperti buku, artikel jurnal, laporan penelitian, atau sumber elektronik yang dapat dipertanggungjawabkan, guna meningkatkan kualitas serta keabsahan argumen. Adapun hasil penelitian skripsi yang telah disetujui pembimbing dan penguji menjadi salah satu penentu kelulusan mahasiswa dalam meraih gelar sarjana.

2.5.1 *N-gram*

N-gram adalah salah satu teknik dalam *text mining* yang digunakan untuk mengukur kemiripan antara dua teks atau dokumen. *N-gram* adalah kumpulan n buah kata yang berurutan dalam sebuah teks. Misalnya, *n-gram* dengan $n=2$ disebut *bigram*, *n-gram* dengan $n=3$ disebut *trigram*, dan seterusnya. Untuk menghitung kemiripan antara dua teks menggunakan *n-gram*, kita dapat menghitung jumlah *n-gram* yang sama di kedua teks dan membaginya dengan jumlah total *n-gram* di kedua teks (Anugrah, 2021). Semakin tinggi nilai kemiripan, semakin mirip kedua teks tersebut.

2.5.2 TF-IDF

Nama lengkap TF-IDF adalah *Term Frequency Inverse Document Frequency*, yaitu metode penghitungan bobot kata dalam sebuah dokumen. TF-IDF dapat digunakan untuk menentukan tingkat kepentingan suatu kata dalam suatu dokumen atau korpus. Rumus yang digunakan untuk menghitung TF-IDF menggunakan persamaan (2.7):

$$tf-idf(t, d) = tf(t, d) \times idf(t) \quad (2.7)$$

Keterangan:

t : term/kata

d : dokumen

TF(t, d): frekuensi kemunculan kata t dalam dokumen d

IDF(t): inverse document frequency dari kata t

Dimana t adalah kata, d adalah dokumen, $tf(t, d)$ adalah frekuensi kata t dalam dokumen d , $tf(t, d)$ dihitung dengan rumus (2.8) :

$$tf(t, d) = \frac{f_{t,d}}{\sum f(w, d)} \quad (2.8)$$

Keterangan:

$f(t,d)$: jumlah kemunculan kata t dalam dokumen d
 $\sum f(w, d)$: total seluruh kata dalam dokumen d
 w : semua kata dalam dokumen

$idf(t)$ adalah frekuensi inversi kata t dalam dokumen, dihitung dengan rumus:

$$IDF(t) = \log \frac{N}{df(t)} \quad (2.9)$$

Keterangan:

N : jumlah seluruh dokumen dalam korpus
 $df(t)$: jumlah dokumen yang mengandung kata t

IDF digunakan untuk mengurangi bobot kata yang terlalu umum. Semakin banyak dokumen yang mengandung kata tersebut, maka nilai IDF akan semakin kecil.

2.5.3 Coherence Score

Kita dapat menggunakan *Coherence Score* dalam *topic mining* untuk mengukur seberapa mudah suatu topik dapat dipahami oleh manusia. Dalam hal ini, *Coherence Score* digunakan sebagai validasi dari hasil *topic mining*. Singkatnya, skor konsistensi mengukur seberapa mirip kata-kata ini satu sama lain. Salah satu dari model evaluasi yang digunakan adalah *Umass* dan *Cv Coherence Score*. *Umass* menghitung seberapa sering dua kata muncul yaitu w_i dan w_j bersamaan pada sebuah *Corpus*.

Umass didefinisikan pada persamaan (2.10):

$$U_{mass}(w_i, w_j) = \sum_{i < j} \log \frac{P(w_i, w_j) + 1}{P(w_i)} \quad (2.10)$$

Keterangan:

w_i, w_j : pasangan kata dalam topik

$P(w_i, w_j)$: jumlah dokumen yang mengandung kedua kata

$P(w_i)$: jumlah dokumen yang mengandung kata w_i

+1: smoothing untuk menghindari $\log(0)$

Coherence UMass mengukur seberapa sering kata-kata dalam satu topik muncul bersama dalam dokumen. Semakin tinggi nilai *coherence*, semakin baik kualitas topik. C_v didefinisikan pada persamaan (2.10):

$$C_v = \frac{\sum_{i,j} \text{sim}(w_i, w_j)}{\text{jumlah pasangan kata}} \quad (2.11)$$

Keterangan:

w_i, w_j : pasangan kata dalam topik

$\text{sim}(w_i, w_j)$: nilai kesamaan (similarity) antar kata

Dimana $P(w_i, w_j)$ adalah banyaknya kemunculan kata w_i dan w_j muncul pada sebuah dokumen. Dan $P(w_i)$, $P(w_j)$ adalah probabilitas kata w_i w_j muncul pada sebuah model latih (Trenquier, 2018).

2.6 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat interpreted, dikembangkan oleh pemrogram asal Belanda Guido van Rossum dan dinamai berdasarkan grup komedi Inggris Monty Python (Rajagopalan, 2021). Sebagai bahasa pemrograman yang berorientasi objek, bersifat skrip, dan interpretatif, Python mengeksekusi kode secara langsung melalui interpreter tanpa memerlukan proses kompilasi (Baliyan et al., 2022).

Sifat interpretatif ini, meskipun membatasi kemampuan optimasi kinerja, membuat *Python* lebih mudah diakses dan sesuai untuk tujuan pendidikan maupun

aplikasi dunia nyata (Nava & Lee, 2023). *Python* telah berkembang menjadi salah satu bahasa pemrograman dengan pertumbuhan tercepat di dunia dan sangat digemari terutama dalam bidang *data science* dan *machine learning* (Rajagopalan, 2021; Sharma et al., 2025)

Bahasa ini menyediakan pustaka bawaan yang luas untuk mendukung berbagai bidang ilmu komputer termasuk kecerdasan buatan, visualisasi data, pengembangan web, serta aplikasi *Internet of Things (IoT)* (Baliyan et al., 2022; Sharma et al., 2025). Paket-paket penting seperti *NumPy*, *SciPy*, *TensorFlow*, *Keras*, *Matplotlib*, dan *SymPy* telah menegaskan peran signifikan *Python* dalam perkembangan industri maupun komputasi ilmiah (Baliyan et al., 2022).

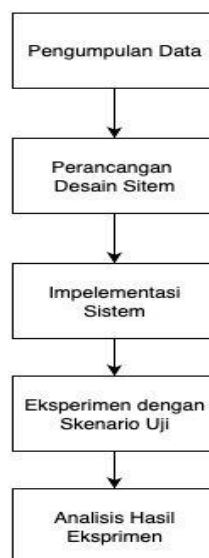
Untuk menjalankan kode *Python*, pengguna hanya perlu menginstal interpreter *Python* dan pustaka terkait pada sistem mereka, yang dapat dijalankan pada berbagai platform seperti Windows, Linux, atau macOS. Dengan ini, *Python* menyediakan platform yang fleksibel dan kuat untuk berbagai jenis komputasi ilmiah, baik numerik maupun simbolik.

BAB III

METODE PENELITIAN

Pembahasan dalam bab ini mencakup desain penelitian, sumber dan karakteristik data yang digunakan, tahapan *text preprocessing*, proses pembentukan *corpus*, implementasi metode *Latent Dirichlet Allocation* (LDA), serta skenario uji coba dan metode evaluasi model. Setiap tahapan dijelaskan secara rinci agar proses penelitian dapat direplikasi dan dipertanggungjawabkan secara ilmiah. Melalui penyusunan metodologi yang terstruktur, penelitian ini diharapkan mampu menghasilkan model topic mining yang optimal dalam mengidentifikasi dan memetakan tren topik skripsi mahasiswa, sehingga memberikan hasil yang valid dan sesuai dengan tujuan penelitian.

3.1 Desain Penelitian



Gambar 3. 1 Desain Penelitian

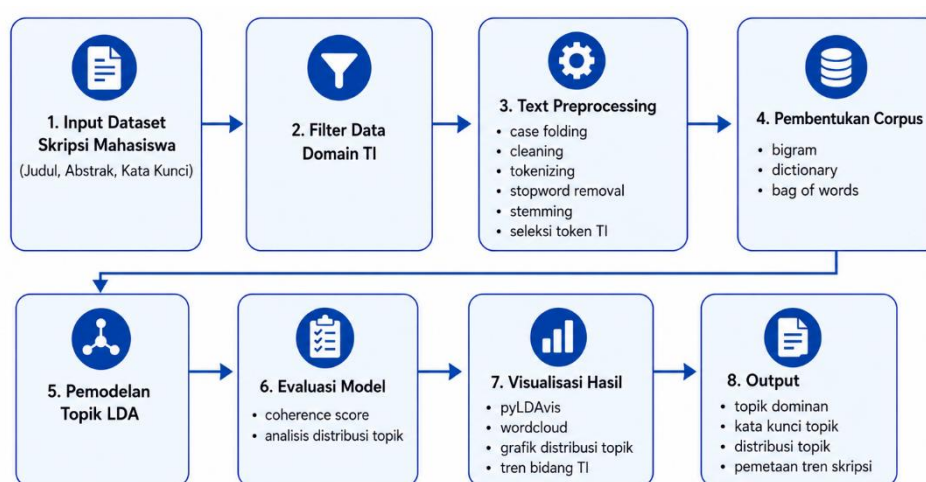
Pada gambar 3.1 desain dan implementasi pada penelitian ini meliputi pengumpulan data kemudian dilanjutkan dengan perancangan desain sistem. Setelah merancang desain sistem, rancangan tersebut diimplementasikan pada sistem agar sistem mendapatkan informasi hasil dari pencocokan fitur. Langkah selanjutnya agar mendapatkan performa terbaik maka diperlukan beberapa eksperimen dari berbagai skenario agar dapat diketahui faktor apa saja yang mempengaruhi kinerja sistem. Setelah mendapatkan hasil dari rangkaian eksperimen dilakukan analisis dari hasil dari rangkaian eksperimen tersebut.

3.1.1 Desain Sistem

Desain sistem pada penelitian ini menggambarkan alur kerja sistem dalam melakukan pemetaan tren topik skripsi mahasiswa menggunakan metode *Latent Dirichlet Allocation* (LDA). Sistem dirancang untuk menerima data skripsi sebagai masukan, kemudian melakukan serangkaian proses pengolahan teks hingga menghasilkan keluaran berupa topik dominan, distribusi topik, visualisasi topik, serta pemetaan tren bidang penelitian. Desain sistem ini bertujuan agar proses penelitian dapat berjalan secara sistematis, mulai dari pengumpulan data, pembersihan teks, pembentukan *corpus*, pemodelan LDA, evaluasi model, sampai tahap interpretasi hasil.

Secara umum, sistem terdiri dari tiga bagian utama, yaitu *input*, *process*, dan *output*. Bagian input berupa dataset skripsi mahasiswa yang diperoleh dari *repository e-theses* UIN Maulana Malik Ibrahim Malang. Data yang digunakan meliputi judul, abstrak, dan kata kunci skripsi. Bagian process mencakup tahapan *text preprocessing*, pembentukan *corpus*, pemodelan topik menggunakan LDA,

evaluasi menggunakan *coherence score*, serta pemetaan hasil topik ke dalam bidang Teknik Informatika. Sementara itu, bagian output menghasilkan informasi berupa jumlah topik optimal, daftar kata dominan setiap topik, topik dominan pada setiap dokumen, visualisasi *pyLDavis*, *wordcloud*, serta grafik distribusi bidang topik penelitian.



Gambar 3.2 Desain Sistem

Desain sistem pada Gambar 3.2 menunjukkan tahapan kerja sistem dalam mengolah data skripsi mahasiswa hingga menghasilkan pemetaan tren topik penelitian. Tahap pertama adalah input data, yaitu dataset skripsi yang terdiri dari judul, abstrak, dan kata kunci. Selanjutnya dilakukan *filter* data domain Teknik Informatika untuk memastikan bahwa dokumen yang digunakan memiliki keterkaitan dengan bidang penelitian yang dianalisis. Setelah itu, data diproses melalui tahap text preprocessing untuk membersihkan teks dan mengubahnya menjadi token yang lebih siap digunakan dalam pemodelan.

Tahap berikutnya adalah pembentukan *corpus*, yaitu proses mengubah dokumen teks menjadi representasi numerik agar dapat diproses oleh model LDA.

Corpus dibentuk melalui tahapan *bigram*, *dictionary*, *filtering* kata, dan representasi *Bag of Words*. Setelah corpus terbentuk, sistem melakukan pemodelan menggunakan LDA untuk mengidentifikasi topik tersembunyi dalam kumpulan dokumen. Model yang dihasilkan kemudian dievaluasi menggunakan *coherence score* untuk melihat kualitas keterkaitan kata dalam topik. Hasil akhir dari sistem berupa topik dominan, kata kunci setiap topik, distribusi topik, visualisasi *pyLDavis*, *wordcloud*, serta pemetaan tren topik skripsi berdasarkan bidang Teknik Informatika.

3.2 Data

Pada penelitian ini, peneliti menggunakan data skripsi yang didapatkan pada tanggal 26 Maret 2026 dari *Etheses* UIN Maulana Malik Ibrahim Malang dan berfokus pada jurusan teknik informatika. Data yang digunakan dalam penelitian ini adalah judul, abstrak, serta kata kunci skripsi yang diperoleh dari *etheses* UIN Maulana Malik Ibrahim Malang menggunakan format CSV. Abstrak-abstrak tersebut berfungsi sebagai kumpulan dokumen yang akan dianalisis untuk menemukan topik-topik yang tersembunyi. Data ini kemudian akan digunakan untuk analisis lebih lanjut menggunakan metode *Latent Dirichlet Allocation* (LDA), yang bertujuan untuk mengidentifikasi topik-topik utama yang terdapat dalam skripsi tersebut. Dengan demikian, data yang dikumpulkan akan membantu dalam proses *topic mining*.

Tabel 3. 1 *Dataset*

Variabel	Definisi Operasional Variabel
Judul	Judul skripsi yang dilakukan masing-masing mahasiswa.
Abstrak	Abstrak merupakan ringkasan singkat dari keseluruhan kandungan dokumen dalam kalimat yang mampu mewakili keseluruhan isi dokumen dengan jelas.
Kata kunci	Kata kunci yang ada pada skripsi mahasiswa.

Tabel 3.1 menjelaskan variabel data yang digunakan dalam penelitian ini. Data yang digunakan terdiri dari tiga atribut utama, yaitu judul, abstrak, dan kata kunci. Judul digunakan untuk mengetahui topik utama dari skripsi mahasiswa secara singkat. Abstrak digunakan karena memuat ringkasan isi penelitian, mulai dari latar belakang, tujuan, metode, hingga hasil penelitian. Sementara itu, kata kunci digunakan untuk membantu mengenali istilah penting yang berkaitan dengan topik skripsi.

3.3 Text Processing

Pada *Text Processing*, data yang telah dikumpulkan akan diproses untuk diolah sebelum dapat digunakan metode LDA. Tahap ini membersihkan dan menstandarisasi teks dari berbagai sumber dokumen. *Text Processing* yang digunakan pada penelitian ini menggunakan *Tokenizing*, *stopword*, *Stemming*, TF-IDF dan n-gram. Pada proses *Tokenizing*, kalimat akan diubah menjadi huruf kecil dengan *lowercasing*.

Tabel 3. 2 Proses *Tokenizing*

Sebelum <i>Tokenizing</i>	Sesudah <i>Tokenizing</i>
Perkembangan penggunaan website saat ini semakin luas, terutama oleh institusi - institusi di Indonesia. Namun untuk setiap website yang di hosting di server, tidak diketahui apakah memiliki kualitas yang baik atau buruk	["perkembangan", "penggunaan", "website", "saat", "ini", "semakin", "luas", ",", "terutama", "oleh", "institusi", "-", "institusi", "di", "Indonesia", ".", "namun", "untuk", "setiap", "website", "yang", "di", "hosting", "di", "server", ",", "tidak", "diketahui", "apakah", "memiliki", "kualitas", "yang", "baik", "atau", "buruk"]

Hasil proses *Tokenizing* ditampilkan pada tabel 3.2 dimana teks yang awalnya merupakan satu paragraf *string* akan dipisah dengan fungsi pada python menjadi

kumpulan kata. Proses yang bersamaan dilakukan fungsi *lowercasing*. Fungsi ini digunakan untuk menghilangkan huruf kapital pada kata.

Teks yang telah dilakukan *Tokenizing* selanjutnya diproses untuk dihilangkan kata hubung yang digunakan dengan *stopword*. Kata-kata yang diproses di *stopword* akan menghilangkan teks yang berulang dan tidak penting untuk memaksimalkan pencarian nantinya.

Tabel 3. 3 Proses *Stopword*

Sebelum Stopword	Sesudah Stopword
["perkembangan", "penggunaan", "website", "saat", "ini", "semakin", "luas", ",", "terutama", "oleh", "institusi", "-", "institusi", "di", "Indonesia", ".", "Namun", "untuk", "setiap", "website", "yang", "di", "hosting", "di", "server", ",", "tidak", "diketahui", "apakah", "memiliki", "kualitas", "yang", "baik", "atau", "buruk"]	["perkembangan", "penggunaan", "website", "luas", "terutama", "institusi", "institusi", "Indonesia", "website", "hosting", "server", "diketahui", "memiliki", "kualitas", "baik", "buruk"]

Stopword dapat menggunakan library NLTK (*natural language toolkit*) untuk menghilangkan kata yang tidak penting dalam teks. Dalam penggunaan NLTK, terdapat kekurangan dimana masih belum bagus dalam mengeliminasi kata dalam Bahasa Indonesia. Hal ini dapat diatasi dengan menggunakan *library* untuk Bahasa Indonesia yaitu *library* sastrawi. Ditampilkan pada tabel 3.3 dimana proses *stopword* menghilangkan kata-kata yang umum digunakan seperti kata setiap, hal, ini, karena, dan sampai. Kata tersebut akan dihapus pada kumpulan kata.

Selanjutnya dilakukan *Stemming* pada *preprocessing*. Proses ini akan mengubah kata menjadi kata dasar atau akar dari kata tersebut.

Tabel 3. 4 Proses *Stemming*

Sebelum <i>Stemming</i>	Sesudah <i>Stemming</i>
["perkembangan", "penggunaan", "website", "luas", "terutama", "institusi", "institusi", "Indonesia", "website", "hosting", "server", "diketahui", "memiliki", "kualitas", "baik", "buruk"]	["kembang", "guna", "website", "luas", "utama", "institusi", "institusi", "Indonesia", "website", "host", "server", "ketahui", "milik", "kualitas", "baik", "buruk"]

Pada tabel 3.4 beberapa kata yang sebelumnya memiliki kata imbuhan seperti kata harinya akan diubah menjadi kata aslinya yaitu hari, lalu pada kata bergerak menjadi kata gerak dan kata bayang-bayang yang merupakan kata berulang diproses menjadi kata bayang.

3.4 Membangun *Corpus*

N-gram yang digunakan nantinya adalah *bigram* yang membuat satu buah kata bisa terdiri dari dua karakter agar tidak merubah makna dari kata tersebut. Selanjutnya dilakukan tf-idf untuk menghitung frekuensi dari kata di sebuah dokumen. Proses pertama yang perlu dilakukan tf-idf adalah menyediakan dokumen terlebih dahulu.

1. D1: Perkembangan penggunaan *website* saat ini semakin luas, terutama oleh institusi - institusi di Indonesia. Namun untuk setiap *website* yang di hosting di server, tidak diketahui apakah memiliki kualitas yang baik atau buruk.
2. D2: Pengujian perangkat lunak memainkan peran yang penting dalam menjaga kualitas perangkat lunak. Akan tetapi, pengujian memakan cukup banyak biaya, waktu, dan tenaga.
3. D3: Terdapat suatu masalah ketika menggunakan banyak animasi untuk satu objek yaitu penampilan karakter yang tidak sesuai dan pergerakan animasi pertarungan yang terlalu cepat

Langkah pertama dalam TF-IDF adalah data diproses untuk *Cleaning*, *stopword*, dan *Stemming* yang selanjutnya di *Tokenizing*.

1. *Document 1*: [perkembangan, penggunaan, *website*, institusi, indonesia, *hosting*, *server*, kualitas, baik, buruk]
2. *Document 2*: [pengujian, perangkat, lunak, memainkan, menjaga, kualitas, memakan, biaya, waktu, tenaga]
3. *Document 3*: [masalah, animasi, objek, penampilan, karakter, pergerakan, pertarungan, cepat]

Langkah kedua, hitung jumlah *Term Frequency* (TF).

TF mewakili frekuensi *term* dalam dokumen. Proses dilakukan dengan menghitung berapa kali setiap *term* muncul di setiap dokumen. Perhitungan ditampilkan pada tabel 3.5:

Tabel 3. 5 *Term Frequency* (TF)

<i>Term</i>	D1	D2	D3
perkembangan	1		
penggunaan	1		
website	1		
institusi	1		
indonesia	1		
hosting	1		
server	1		
kualitas	1	1	
baik	1		
buruk	1		
pengujian		1	
perangkat		1	
lunak		1	
memainkan		1	
menjaga		1	
memakan		1	
biaya		1	
waktu		1	
tenaga		1	
masalah			1
animasi			1
objek			1
penampilan			1
karakter			1

pergerakan			1
pertarungan			1
cepat			1

Proses penentuan *term frequency* ditampilkan pada tabel 3.5 dimana terdapat dua dokumen yang digunakan. Kedua dokumen D0, D1 memiliki kata kualitas didalamnya. Selanjutnya proses normalisasi $tf(t,d)$ dengan menggunakan rumus (2.2),

$$tf("kondisi", D0) = \frac{\text{banyak term kondisi muncul pada dokun}}{\text{total term yang muncul pada dokumei}}$$

Sebagai contoh, pada dokumen D1 terdapat token hasil *preprocessing* sebagai berikut:

[perkembangan, penggunaan, website, institusi, indonesia, hosting, server, kualitas, baik, buruk]

Jumlah seluruh *term* pada D1 adalah 10. Apabila dihitung nilai TF untuk *term* “perkembangan”, maka *term* tersebut muncul sebanyak 1 kali pada D1. Dengan demikian, nilai TF *term* “perkembangan” pada D1 adalah:

$$tf(\text{perkembangan}, D_1) = \frac{1}{10} = 0,1$$

Hasil tersebut menunjukkan bahwa *term* “perkembangan” memiliki bobot frekuensi sebesar 0,1 pada dokumen D1. Nilai ini diperoleh karena *term* tersebut hanya muncul satu kali dari total 10 *term* yang terdapat pada dokumen D1. sehingga diperoleh hasil tf dengan nilai 0,1.

Selain itu, pada dokumen D2 terdapat token sebagai berikut:

[pengujian, perangkat, lunak, memainkan, menjaga, kualitas, memakan, biaya, waktu, tenaga]

Jumlah seluruh *term* pada D2 juga sebanyak 10. Apabila dihitung nilai TF untuk *term* “kualitas”, maka *term* tersebut muncul sebanyak 1 kali pada D2. Dengan demikian, nilai TF *term* “kualitas” pada D2 adalah:

$$tf(kualitas, D_2) = \frac{1}{10} = 0,1$$

Sementara itu, pada dokumen D3 terdapat token sebagai berikut:

[masalah, animasi, objek, penampilan, karakter, pergerakan, pertarungan, cepat]

Jumlah seluruh *term* pada D3 adalah 8. Apabila dihitung nilai TF untuk *term* “animasi”, maka *term* tersebut muncul sebanyak 1 kali pada D3. Dengan demikian, nilai TF *term* “animasi” pada D3 adalah:

$$tf(animasi, D_3) = \frac{1}{8} = 0,125$$

Hasil normalisasi *term frequency* selanjutnya ditampilkan pada tabel 3.6.

Tabel 3. 6 Normalisasi Term Frequency (TF)

<i>Term</i>	D0	D1	D2
perkembangan	0,1		
Penggunaan	0,1		
Website	0,1		
Intitusi	0,1		
Indonesia	0,1		
hosting	0,1		
Server	0,1		
Kualitas	0,1	0,1	
Baik	0,1		
Buruk	0,1		
Pengujian		0,1	
Perangkat		0,1	
Lunak		0,1	
Memainkan		0,1	
Menjaga		0,1	
Memakan		0,1	
Baiya		0,1	
Waktu		0,1	
Tenaga		0,1	
Masalah			0,125
Animasi			0,125
Objek			0,125
Penampilan			0,125

Karakter			0,125
Pergerakan			0,125
Pertarungan			0,125
cepat			0,125

Tabel 3.6 menampilkan hasil dari normalisasi TF yang telah dilakukan pada tiga dokumen sebelumnya. Terdapat beberapa hasil yang berbeda pada D2, memiliki nilai TF yaitu 0,125 karena jumlah kemunculan *term* pada dokumen adalah 8.

Selanjutnya dilakukan perhitungan untuk mencari IDF. Menggunakan perhitungan rumus (2.3), perhitungan $idf(t)$, perlu mencari frekuensi kemunculan *term* pada semua dokumen atau korpus $df(t)$. Seperti pada kata kualitas dimana memiliki $df(t)$ berjumlah 2.

$$idf("kondisi") = \left(\frac{3}{1}\right) = 0.477$$

Selanjutnya dilakukan pencarian tf-idf dengan formula (2.7), didapatkan hasil:

$$tf - idf("kondisi, d_0) = 0.1 \times 0.477 = 0.0477$$

Hasil tersebut akan digunakan sebagai bobot kata kondisi pada dokumen D0.

Proses TF=IDF ditampilkan pada tabel 3.7.

Tabel 3. 7 TF-IDF

TERM	DF	IDF	TF-IDF		
			D0	D1	D2
Perkembangan	1	0,477	0,0477	0	0
penggunaan	1	0,477	0,0477	0	0
Website	1	0,477	0,0477	0	0
Institusi	1	0,477	0,0477	0	0
Indonesia	1	0,477	0,0477	0	0
Hosting	1	0,477	0,0477	0	0
Server	1	0,477	0,0477	0	0
kualitas	2	0,176	0,0176	0,0176	0
Baik	1	0,477	0,0477	0	0
Buruk	1	0,477	0,0477	0	0
Pengujian	1	0,477	0	0,0477	0
perangkat	1	0,477	0	0,0477	0
Lunak	1	0,477	0	0,0477	0

memainkan	1	0,477	0	0,0477	0
menjaga	1	0,477	0	0,0477	0
Memakan	1	0,477		0,0477	
Biaya	1	0,477		0,0477	
Waktu	1	0,477		0,0477	
Tenaga	1	0,477		0,0477	
Masalah	1	0,477			0,0596
animasi	1	0,477			0,0596
Objek	1	0,477			0,0596
Penampilan	1	0,477			0,0596
Karakter	1	0,477			0,0596
Pergerakan	1	0,477			0,0596
Pertarungan	1	0,477			0,0596
Cepat	1	0,477			0,0596

Tabel 3.7 memberikan hasil dari perkalian antara nilai TF dan IDF, sehingga menghasilkan bobot (TF-IDF) untuk setiap *term* pada masing-masing dokumen. Bobot yang diperoleh mencerminkan seberapa penting suatu *term* terhadap dokumen tersebut. Pada dokumen D0 diperoleh bobot sebesar 0,0477 untuk hampir semua *term* karena setiap *term* hanya muncul satu kali pada dokumen yang berisi total 10 kata, dan masing-masing memiliki nilai IDF sebesar 0,477. *Term* kualitas memiliki bobot lebih kecil, yaitu 0,0176 pada D0 dan D1 karena *term* ini muncul pada dua dokumen sekaligus, sehingga nilai IDF-nya lebih rendah (0,176). Pada dokumen D1 diperoleh pola yang sama, dengan bobot 0,0477 untuk semua *term* yang hanya muncul pada D1, sementara *term* kualitas tetap bernilai 0,0176. Berbeda dengan dokumen D2, karena total kata hanya berjumlah 8, maka nilai TF untuk setiap *term* menjadi lebih besar (0,125). Hal ini membuat nilai bobot TF-IDF untuk semua *term* di D2 menjadi 0,0596, lebih tinggi dibandingkan *term* pada D0 maupun D1. Dengan demikian, perbedaan bobot antar dokumen dipengaruhi oleh dua faktor, yaitu jumlah kemunculan *term* dalam dokumen (TF) dan jumlah dokumen yang mengandung *term* tersebut (DF yang memengaruhi

IDF). Semakin jarang *term* muncul di seluruh koleksi dokumen, semakin besar nilai IDF dan bobot TF-IDF-nya.

3.5 Implementasi Metode

Setelah data telah melewati tahap *Text Processing*, tahap selanjutnya adalah pembuatan model LDA. Proses LDA pada gensim dapat diproses seperti pada rumus untuk mendapatkan persamaan (2.1) dengan penentuan parameter dirichlet yaitu α dan β diikuti dengan penentuan besar topik atau K . Pada sampel selanjutnya akan digunakan parameter dengan nilai $\alpha = 0.1$, $\beta = 0.1$, $K = 3$.

Kumpulan dokumen yang telah di tokenisasi akan diubah dalam bentuk korpus yang selanjutnya diinisialisasi topik untuk semua kata dalam dokumen secara random (topik 1 adalah K_1 , dst). Berikut adalah simulasi untuk 3 dokumen (D1, D2, dan D3) dan topik tersembunyi.

Tabel 3. 8 Insialisasi Topik

No	D1	K	D2	K	D3	K
1	Perkembangan	2	Pengujian	3	masalah	1
2	Penggunaan	1	Perangkat	2	Animasi	2
3	Website	3	Lunak	1	Objek	1
4	Institusi	2	Kualitas	2	Penampilan	1
5	Indonesia	1	Biaya	3	Karakter	2
6	Server	3	Waktu	3	Pergerakan	1
7	Kualitas	1	Tenaga	3	Pertarungan	1
8	Hosting	2	Penting	1	Cepat	1
9	Baik	3	Menjaga	1	Menggunakan	2
10	buruk	1	peran	2	sesuai	3

Tabel 3.8 dilakukan pemberian topik secara random pada teks yang telah diproses pada setiap dokumen. Karena pada sebelumnya telah ditentukan topik sebanyak 3, maka pada setiap kata akan menjadi kata pada topik satu, dua dan tiga. Dari tabel 3.8 dapat diketahui distribusi topik terhadap kata dan distribusi dokumen terhadap topik yang ditampilkan pada tabel 3.9 dan 3.10.

Tabel 3. 9 Distribusi Topik Pada Kata

<i>Term</i>	K1	K2	K3
Perkembangan	0	1	0
Penggunaan	1	0	0
Website	0	0	1
Institusi	0	1	0
Indonesia	1	0	0
Sever	0	0	1
kualitas	1	1	0
Hositng	0	1	0
Baik	0	0	1
Buruk	1	0	0
Pengujian	1	0	0
Perangkat	0	0	1
Lunak	1	0	0
Biaya	0	0	1
Waktu	0	1	0
Tenaga	1	0	0
Penting	0	1	0
Menjaga	1	0	0
Peran	0	0	1
Masalah	0	0	1
Animasi	0	1	0
Objek	1	0	0
penampilan	0	0	1
Karakter	0	1	0
Pergerakan	1	0	0
Pertarungan	0	1	0
Cepat	0	0	1
Menggunakan	1	0	0
sesuai	0	1	0

Persebaran kata terhadap topik ditampilkan pada tabel 3.9 dimana kata pada topik K1 memiliki kata bank cukup banyak dibandingkan pada K2 dan K3, kata ekonomi memiliki penempatan pada topik K1 sebanyak tiga kali, lalu dilanjutkan pada K2 sebanyak dua kali dan K3 sebanyak sekali. Pada kata investasi, mendapat penempatan yang sama pada setiap topik. Kata krisis yang muncul hanya satu kali pada dokumen D2 saja mendapat asumsi sebagai topik K2. Kata minyak memiliki penempatan pada K2 terbesar sebanyak 4 kali disbandingkan pada K1 dan K3. Kata resesi memiliki penempatan sama pada K1 dan K2 sebanyak dua kali dan hanya sekali pada K3. Kata usia memiliki penempatan sekali pada K1 dan K3 dan bernilai

0 karena hanya dua kali kata usia muncul pada semua dokumen.

Tabel 3. 10 Distribusi Dokumen Terhadap Topik

	D1	D2	D3
K1	3	4	3
K2	3	3	4
K3	4	3	3

Tabel 3.10 menampilkan distribusi dokumen terhadap topik yang telah ditetapkan secara random sebelumnya. Terlihat pada dokumen D1 memiliki persebaran topik K1 dan K2 sebanyak 3 dan K3 sebanyak 4. Pada dokumen D2 memiliki persebaran topik K1 sebanyak 4, K2 dan K3 sebanyak 3 kata. Pada dokumen D3 memiliki persebaran topik K1 dan K3 sebanyak 3, K3 sebanyak 4.

Selanjutnya menggunakan persamaan (2.5) dan (2.6) untuk mendapat probabilitas kata terhadap topik:

$$P(Z = K|z, w, d) = n_{d,k} + \alpha \times \frac{m_{w,k} + \beta}{\sum_{i=1}^V m_{i,k} + W\beta}$$

Pada simulasi ini, akan di lakukan sample kata pertama pada dokmen 1 yaitu “perkembangan”.

$$P(Z = K_1 | z, perkembangan, d_1) = (3 + 0.1) \times \frac{1 + 0.1}{10 + 28 \times 0.1} = 0.0320$$

$$P(Z = K_2 | z, perkembangan, d_1) = (3 + 0.1) \times \frac{1 + 0.1}{9 + 28 \times 0.1} = 0.2890$$

$$P(Z = K_3 | z, perkembangan, d_1) = (4 + 0.1) \times \frac{1 + 0.1}{11 + 28 \times 0.1} = 0.0225$$

Dari hasil diatas dapat dilihat bahwa topik 2 memiliki hasil yang lebih besar, sehingga kata “perkembangan” pada dokumen 1 merupakan topik 2.

Tabel 3. 11 Kata Z Setelah Gibbs Sampling Pada D1

No.	D1	K	D2	K	D3	K
1	Perkembangan	1	Pengujian	1	Masalah	3
2	Penggunaan	1	Perangkat	3	Animasi	2
3	Website	3	Lunak	1	Objek	1

4	Institusi	2	Kualitas	2	penampilan	3
5	Indonesia	1	Biaya	3	Karakter	2
6	server	3	Waktu	2	Pergerakan	1
7	Kualitas	1	Tenaga	1	Pertarungan	2
8	Hosting	2	Pentign	2	Cepat	3
9	Baik	3	Menjaga	1	Menggunakan	1
10	buruk	1	peran	3	sesuai	2

Pada tabel 3.11, pada kata perkembangan di dokumen D1 mengalami perubahan pada penempatan topik yang sebelumnya dilakukan secara random yaitu sebagai topik 1, kini setelah diproses menggunakan *gibbs sampling* menjadi topik 2. Proses distribusi topik terhadap kata dan dokumen terhadap topik dilakukan lagi menyesuaikan data yang telah diperbarui.

Tabel 3. 12 Hasil Penyesuaian Distribusi Topik Terhadap Kata

Kata	K1	K2	K3
Perkembangan	1	0	0
Penggunaan	1	0	0
Website	0	0	1
Institusi	0	1	0
Indonesia	1	0	0
Sever	0	0	1
kualitas	2	1	0
Hositng	0	1	0
Baik	0	0	1
Buruk	1	0	0
Pengujian	1	0	0
Perangkat	0	0	1
Lunak	1	0	0
Biaya	0	0	1
Waktu	0	1	0
Tenaga	1	0	0
Penting	0	1	0
Menjaga	1	0	0
Peran	0	0	1
Masalah	0	0	1
Animasi	0	1	0
Objek	1	0	0
penampilan	0	0	1
Karakter	0	1	0
Pergerakan	1	0	0
Pertarungan	0	1	0
Cepat	0	0	1
Menggunakan	1	0	0
sesuai	0	1	0

Tabel 3.12 merupakan penyesuaian dari distribusi yang dilakukan sebelumnya dalam topik terhadap kata. Kata perkembangan pada dokumen D1 yang telah ditetapkan menjadi topik 2, sehingga kata perkembangan pada topik K1 berkurang dan pada topik K2 bertambah. Selanjutnya penyesesuaian distribusi dokumen terhadap topik ditampilkan pada tabel 3.13:

Tabel 3. 13 Hasil Penyesuaian Distribusi Dokumen Terhadap Topik

	D1	D2	D3
K1	5	3	3
K2	2	2	4
K3	3	5	3

Tabel 3.13 menyesuaikan perubahan yang telah terjadi pada kata perkembangan, kini pada dokumen D1, mengalami perubahan Dimana kata perkembangan sebelumnya mermerupakan topik K1 kini menjadi K2, sehingga K1 pada dokumen berkurang satu dan K2 pada dokumen D1 bertambah satu. Proses pada diatas akan dilakukan terus menerus hingga semua semua kata pada *Corpus* sudah dilakukan proses *gibbs sampling* dan mendapatkan topik yang sesuai berdsarkan probabilitas topik LDA.

Metode LDA pada penelitian ini akan menggunakan *library* python yaitu gensim. Pemanggilan LDA dengan *library* gensim berisi parameter berupa *id2word* yang berasal dari Kumpulan kata yang di map dengan ID, parameter *Corpus* yang merupakan kumpulan ID dengan jumlah yang muncul pada setiap dokumen, parmater *num_topic* untuk menentukan berapa banyak topik yang akan diekstrak dalam model pelatihan, Output yang dilkeuarkan dari LDA berupa Kumpulan topik yang akan ditampilkan dengan grafik dan *wordcloud*.

3.6 Skenario Uji Coba

Tahap skenario uji coba dilakukan untuk mengetahui pengaruh dari variasi proses text *preprocessing* dan parameter model terhadap hasil analisis topik menggunakan metode *Latent Dirichlet Allocation* (LDA). Skenario uji coba ini dirancang agar diperoleh konfigurasi model yang menghasilkan nilai *Coherence Score* terbaik serta topik yang paling relevan dan interpretatif terhadap kumpulan data skripsi dari etheases UIN Maulana Malik Ibrahim Malang.

Uji coba dilakukan dengan tiga skenario berbeda yang disesuaikan dengan tahapan implementasi pada sistem, yaitu filtering kata pada *Corpus*, penggunaan bigram, serta penentuan jumlah topik optimal. Evaluasi dilakukan menggunakan metrik *Coherence Score* (c_v) dan visualisasi word cloud untuk melihat dominasi kata pada setiap topik.

Skenario pertama bertujuan untuk mengetahui pengaruh proses filtering kata pada tahap pembentukan *Dictionary* terhadap kualitas topik yang dihasilkan oleh model LDA. Pada implementasi yang telah dilakukan, filtering kata dilakukan menggunakan parameter $no_below = 5$ dan $no_above = 0.5$, yaitu menghapus kata yang terlalu jarang muncul dan kata yang terlalu sering muncul dalam *Corpus*.

Pada skenario ini digunakan dua pendekatan, yaitu:

- a) tanpa filtering kata pada *Dictionary*
- b) dengan filtering kata menggunakan parameter no_below dan no_above

Parameter model LDA yang digunakan adalah $\alpha = 0.1$, $\beta = 0.1$, jumlah topik (K) = 3, dan $passes = 10$ sesuai dengan implementasi pada kode program.

Evaluasi dilakukan dengan membandingkan nilai *Coherence Score* serta kualitas interpretasi topik yang dihasilkan. Proses filtering ini bertujuan untuk mengurangi noise dalam *Corpus*, karena kata yang terlalu jarang cenderung tidak stabil, sedangkan kata yang terlalu sering muncul tidak mampu membedakan topik secara signifikan.

Pendekatan ini sejalan dengan konsep *preprocessing* dalam topic modeling yang menekankan pentingnya penyaringan kata untuk meningkatkan kualitas model. Penelitian oleh Alattar (2021) menunjukkan bahwa pendekatan *Filtered-LDA* mampu meningkatkan kualitas topik dengan mengurangi data yang tidak relevan, sedangkan Zimmermann (2024) menyatakan bahwa filtering merupakan faktor penting dalam meningkatkan *Coherence Score* pada model LDA. Oleh karena itu, hasil dari skenario ini diharapkan menunjukkan bahwa penggunaan filtering kata menghasilkan topik yang lebih jelas dan representatif.

Skenario kedua bertujuan untuk mengetahui pengaruh penggunaan bigram terhadap kualitas topik yang dihasilkan oleh model LDA. Pada tahap pembentukan *Corpus* dalam penelitian ini, dilakukan pembentukan bigram menggunakan library Gensim untuk menggabungkan dua kata yang sering muncul secara berurutan, seperti “machine learning” atau “sistem informasi”.

Pada skenario ini digunakan dua pendekatan, yaitu:

- a) menggunakan unigram (tanpa bigram)
- b) menggunakan bigram

Parameter model LDA tetap digunakan sama, yaitu $\alpha = 0.1$, $\beta = 0.1$, jumlah topik (K) = 3, dan passes = 10.

Evaluasi dilakukan dengan membandingkan nilai *Coherence Score* dan keterbacaan kata dominan pada setiap topik. Penggunaan bigram diharapkan mampu meningkatkan kualitas representasi teks karena mampu menangkap konteks antar kata yang tidak dapat ditangkap oleh unigram.

Penelitian oleh Chakkarwar dan Tamane (2023) menunjukkan bahwa penggunaan bigram dalam LDA mampu meningkatkan nilai *coherence* dan menghasilkan topik yang lebih kontekstual. Selain itu, Griffiths et al. (2007) menjelaskan bahwa penggunaan collocation seperti bigram dapat meningkatkan kualitas topic model karena mampu merepresentasikan hubungan antar kata dalam suatu konteks. Oleh karena itu, hasil dari skenario ini diharapkan menunjukkan bahwa penggunaan bigram menghasilkan topik yang lebih bermakna dan mudah diinterpretasikan.

Skenario ketiga bertujuan untuk menentukan jumlah topik optimal yang menghasilkan model terbaik berdasarkan nilai *Coherence Score*. Pada implementasi dalam kode program, dilakukan pengujian jumlah topik dari $K = 2$ hingga $K = 10$, kemudian dipilih nilai K dengan *Coherence Score* tertinggi sebagai model terbaik.

Parameter model yang digunakan tetap sama, yaitu $\alpha = 0.1$, $\beta = 0.1$, dan $\text{passes} = 10$. Setiap model LDA dilatih menggunakan *Corpus* yang sama, kemudian dihitung nilai *Coherence Score* untuk setiap variasi jumlah topik.

Evaluasi dilakukan dengan membandingkan nilai *Coherence Score* serta menganalisis keterbacaan kata dominan pada setiap topik. Berdasarkan hasil implementasi, diperoleh nilai *coherence* tertinggi sebesar 0.5581 pada jumlah topik

$K = 6$, sehingga nilai tersebut dipilih sebagai jumlah topik optimal dalam penelitian ini .

Penentuan jumlah topik merupakan faktor penting dalam topic modeling karena jumlah topik yang terlalu sedikit akan menghasilkan topik yang terlalu umum, sedangkan jumlah topik yang terlalu banyak dapat menyebabkan tumpang tindih antar topik. Selain itu, *Coherence Score* digunakan sebagai metrik evaluasi karena mampu mengukur keterkaitan semantik antar kata dalam topik secara kuantitatif (Röder et al., 2015). Oleh karena itu, skenario ini digunakan untuk menentukan konfigurasi model LDA terbaik yang akan digunakan pada tahap analisis hasil.

Ketiga skenario uji coba di atas disusun berdasarkan tahapan implementasi pada sistem, yaitu *preprocessing*, pembentukan *Corpus*, dan pemodelan LDA. Hasil dari setiap skenario dibandingkan untuk memperoleh konfigurasi terbaik yang menghasilkan nilai *Coherence Score* tertinggi serta topik yang paling relevan terhadap kumpulan data skripsi.

3.7 Evaluasi

Tahap evaluasi dilakukan untuk menilai kualitas model *Latent Dirichlet Allocation* (LDA) yang dihasilkan berdasarkan proses pemodelan yang telah dilakukan. Evaluasi ini bertujuan untuk menentukan konfigurasi model terbaik yang mampu menghasilkan topik dengan kualitas paling baik dan mudah diinterpretasikan.

Dalam penelitian ini, evaluasi model dilakukan menggunakan metrik *Coherence Score* (c_v). *Coherence Score* digunakan untuk mengukur tingkat

keterkaitan semantik antar kata dalam satu topik. Semakin tinggi nilai *coherence* , maka semakin baik kualitas topik yang dihasilkan karena kata-kata dalam topik tersebut memiliki hubungan makna yang kuat.

Proses evaluasi dilakukan secara bertahap sesuai dengan implementasi pada kode program, yaitu dengan menguji beberapa variasi jumlah topik (K). Pada penelitian ini, jumlah topik diuji dalam rentang $K = 2$ hingga $K = 10$. Untuk setiap nilai K, dilakukan pelatihan model LDA menggunakan *Corpus* dan *Dictionary* yang telah dibentuk pada tahap sebelumnya.

Setelah model dilatih, dilakukan perhitungan *Coherence Score* menggunakan metode *c_v*. Perhitungan ini memanfaatkan dokumen hasil *preprocessing*, *Dictionary*, serta model LDA yang telah terbentuk. Nilai *coherence* yang dihasilkan kemudian disimpan untuk setiap variasi jumlah topik.

Selanjutnya, dilakukan proses pemilihan model terbaik dengan cara membandingkan seluruh nilai *Coherence Score* yang diperoleh. Model dengan nilai *coherence* tertinggi dipilih sebagai model optimal. Berdasarkan hasil pengujian yang telah dilakukan, diperoleh nilai *coherence* tertinggi sebesar **0,5581** pada jumlah topik $K = 6$, sehingga model dengan 6 topik ditetapkan sebagai model terbaik dalam penelitian ini .

Selain evaluasi secara kuantitatif menggunakan *Coherence Score*, dilakukan juga evaluasi secara kualitatif dengan menganalisis kata-kata dominan pada setiap topik. Analisis ini bertujuan untuk memastikan bahwa topik yang dihasilkan memiliki makna yang jelas, tidak tumpang tindih, dan sesuai dengan konteks data skripsi yang dianalisis.

Coherence Score dipilih sebagai metrik evaluasi karena mampu memberikan ukuran yang objektif terhadap kualitas topik. Menurut Röder et al. (2015), *Coherence Score* merupakan salah satu metode evaluasi yang efektif dalam topic modeling karena mempertimbangkan keterkaitan antar kata dalam suatu topik berdasarkan distribusi kemunculan kata pada *Corpus*.

BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi metode yang telah dirancang pada bab sebelumnya, serta analisis terhadap hasil yang diperoleh. Proses yang dilakukan meliputi pengolahan data, pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA), evaluasi model, serta visualisasi hasil untuk memberikan pemahaman yang lebih komprehensif terhadap pola topik yang terbentuk.

Pembahasan dalam bab ini mencakup tahapan *text preprocessing*, pembentukan *corpus*, proses pemodelan LDA, evaluasi model menggunakan *coherence score* dan distribusi topik, serta visualisasi hasil menggunakan *pyLDavis*, grafik distribusi topik, dan *wordcloud*. Selain itu, dilakukan analisis mendalam terhadap dominasi topik, hubungan antar topik, serta tren penelitian mahasiswa berdasarkan hasil pemodelan.

Melalui analisis yang dilakukan, bab ini bertujuan untuk menginterpretasikan hasil penelitian secara ilmiah, sehingga dapat memberikan gambaran mengenai struktur topik, kecenderungan penelitian mahasiswa, serta kualitas model yang dihasilkan.

4.1 Data Penelitian

Pada penelitian ini, data yang digunakan merupakan data skripsi yang diperoleh dari repository eteses UIN Maulana Malik Ibrahim Malang pada jurusan Teknik Informatika dalam rentang tahun 2020 - 2025 dan didapatkan menggunakan

format CSV. Dataset memuat beberapa atribut utama, yaitu judul, abstrak, dan kata kunci. Dalam implementasi pemodelan, atribut yang digunakan sebagai sumber utama analisis adalah abstrak, karena abstrak memuat ringkasan isi penelitian yang lebih representatif dibandingkan judul.

Tabel 4. 1 Informasi Dataset

Komponen	Nilai
Jumlah data	3961 dokumen
Jumlah kolom	7
Format	CSV
Kolom utama	Abstrak ID

Tabel 4.1 menunjukkan gambaran umum dataset yang digunakan dalam penelitian. Jumlah data sebanyak 3961 dokumen menunjukkan bahwa dataset memiliki ukuran yang cukup besar untuk analisis topic modeling. Jumlah kolom sebanyak 7 menunjukkan bahwa dataset memiliki atribut yang lengkap. Kolom utama yang digunakan adalah Abstrak ID karena mampu merepresentasikan isi penelitian secara lebih komprehensif dibandingkan kolom lainnya.

Tabel 4. 2 Nama Kolom Dataset

No	Kolom
1	ID
2	Penulis
3	Tahun
4	Judul
5	<i>Supervisor</i>
6	Abstrak ID
7	<i>Keywords</i>

Tabel 4.2 menjelaskan struktur atribut dalam dataset. Setiap kolom memiliki fungsi yang berbeda, di mana ID digunakan sebagai identitas unik, Penulis dan *Supervisor* sebagai metadata, Tahun untuk informasi temporal, Judul sebagai gambaran singkat penelitian, dan *Keywords* sebagai pendukung analisis. Kolom

Abstrak ID menjadi fokus utama karena berisi ringkasan isi penelitian yang paling relevan untuk proses *topic mining*.

Tabel 4. 3 Distribusi Data Berdasarkan Tahun

Tahun	Jumlah
2020	614
2021	617
2022	813
2023	853
2024	970
2025	93

Tabel 4.3 menunjukkan distribusi jumlah skripsi berdasarkan tahun. Terlihat bahwa jumlah skripsi mengalami peningkatan dari tahun ke tahun dengan puncaknya pada tahun 2024. Hal ini menunjukkan adanya peningkatan aktivitas penelitian mahasiswa. Sementara itu, jumlah data pada tahun 2025 relatif kecil karena kemungkinan data belum lengkap atau masih dalam proses pengumpulan.

Tabel 4. 4 Statistik Panjang Abstrak

Parameter	Nilai
Jumlah data valid	3911
Rata-rata	504 kata
Minimum	1 kata
Maksimum	2132 kata

Tabel 4.4 menunjukkan statistik panjang abstrak dalam dataset. Rata-rata panjang abstrak sebesar 504 kata menunjukkan bahwa dokumen memiliki informasi yang cukup kaya untuk dianalisis. Namun, terdapat variasi yang cukup besar antara panjang minimum dan maksimum, yang menunjukkan bahwa tidak semua dokumen memiliki kualitas atau kelengkapan yang sama. Variasi ini dapat memengaruhi distribusi kata dalam proses *modeling*.

Tabel 4. 5 Contoh Data Awal

ID	Penulis	Tahun	Judul
18282	Adiyono	2020	Approximate String Matching
21010	Ahmad	2020	Box Muller Gaussian

ID	Penulis	Tahun	Judul
24115	Fadhil	2020	Algoritma Genetika
19840	Islam Nur	2020	Identifikasi Lahan
18153	Mauludy	2020	Naive Bayes

Tabel ini menampilkan contoh data awal dari dataset untuk memberikan gambaran nyata mengenai isi data yang digunakan. Dari tabel tersebut terlihat bahwa topik penelitian sangat beragam, mulai dari algoritma, kecerdasan buatan, hingga pengolahan citra. Variasi ini menunjukkan bahwa dataset memiliki cakupan topik yang luas, sehingga sangat cocok untuk dilakukan analisis *topic mining*. Dataset memiliki jumlah yang besar, struktur yang lengkap, serta variasi topik yang luas. Hal ini menjadikan dataset sangat sesuai untuk analisis menggunakan metode LDA. Namun, ditemukan adanya variasi panjang dokumen dan campuran bahasa yang berpotensi menjadi noise, sehingga diperlukan tahap *preprocessing* untuk meningkatkan kualitas data

4.1.1 Seleksi Data

Pada tahap awal, dataset yang digunakan dalam penelitian ini berjumlah 3961 dokumen. Data tersebut diperoleh dari repository e-theses UIN Maulana Malik Ibrahim Malang dan terdiri dari beberapa atribut utama, yaitu judul, abstrak, dan kata kunci. Ketiga atribut tersebut digunakan sebagai sumber teks utama karena mampu merepresentasikan isi penelitian dari setiap dokumen skripsi. Sebelum digunakan dalam proses pemodelan Latent Dirichlet Allocation (LDA), data terlebih dahulu melalui proses seleksi agar dokumen yang dianalisis benar-benar memiliki keterkaitan dengan bidang Teknik Informatika.

Tahap pertama adalah filter awal domain Teknik Informatika. Pada tahap ini, setiap dokumen diperiksa berdasarkan keberadaan istilah yang berkaitan dengan bidang Teknik Informatika pada gabungan judul, kata kunci, dan abstrak. Berdasarkan hasil proses tersebut, sebanyak 3555 dokumen atau 89,75% dari keseluruhan dataset dinyatakan lolos filter awal TI. Sementara itu, sebanyak 406 dokumen atau 10,25% dinyatakan gugur karena tidak ditemukan sinyal istilah Teknik Informatika yang cukup pada dokumen tersebut.

Setelah filter awal dilakukan, dokumen yang lolos kemudian melalui proses seleksi token TI untuk pembentukan corpus LDA. Pada tahap ini, token yang dipertahankan adalah token yang berkaitan dengan istilah Teknik Informatika, seperti *machine learning*, *data mining*, *web*, *mobile*, *IoT*, *sensor*, *database*, *Software Engineering*, *game*, serta istilah teknis lainnya. Berdasarkan hasil seleksi token TI, sebanyak 3349 dokumen atau 84,55% dari dataset awal dinyatakan valid untuk digunakan dalam pembentukan corpus. Sementara itu, sebanyak 206 dokumen atau 5,20% gugur karena jumlah token istilah TI yang dimiliki tidak memenuhi batas minimal yang telah ditentukan.

Pada tahap berikutnya dilakukan pembentukan dictionary dan corpus. Hasil proses menunjukkan bahwa seluruh dokumen yang valid pada tahap seleksi token TI tetap memiliki representasi corpus yang dapat digunakan, sehingga jumlah valid corpus final LDA tetap sebanyak 3349 dokumen atau 84,55%. Tidak terdapat dokumen yang gugur karena corpus kosong. Dengan demikian, sebanyak 3349 dokumen digunakan sebagai data final dalam proses pemodelan LDA.

Setelah pemodelan dan pemetaan bidang dilakukan, sebanyak 2090 dokumen atau 52,76% dari dataset awal berhasil dipetakan ke dalam enam bidang Teknik Informatika, yaitu Multimedia & Game, Sistem Informasi, Intelligence System, IoT & Robotic, *Software Engineering*, dan Web & Mobile Programming. Sementara itu, sebanyak 1259 dokumen atau 31,78% tidak terklasifikasi ke dalam enam bidang tersebut. Dokumen tidak terklasifikasi bukan berarti tidak digunakan dalam proses LDA, melainkan tidak memiliki kecocokan yang cukup kuat terhadap kamus pemetaan enam bidang TI yang digunakan pada tahap interpretasi bidang.

Tabel 4. 6 Rincian Seleksi Data

Tahap	Jumlah Dokumen	Persentase dari Dataset Awal
Dataset awal	3961	100,00%
Lolos filter awal TI	3555	89,75%
Gugur filter awal TI	406	10,25%
Valid seleksi token TI	3349	84,55%
Gugur seleksi token TI	206	5,20%
Valid corpus final LDA	3349	84,55%
Gugur corpus kosong	0	0,00%
Berhasil dipetakan ke 6 bidang TI	2090	52,76%
Tidak terklasifikasi bidang	1259	31,78%

Berdasarkan Tabel 4.6, dapat diketahui bahwa proses seleksi data dilakukan secara bertahap dan terukur. Dari total 3961 dokumen, sebanyak 3555 dokumen lolos pada filter awal domain Teknik Informatika. Hal ini menunjukkan bahwa sebagian besar dokumen dalam dataset memiliki keterkaitan dengan bidang TI. Namun, terdapat 406 dokumen yang gugur karena tidak ditemukan sinyal istilah Teknik Informatika pada judul, kata kunci, maupun abstrak.

Pada tahap seleksi token TI, jumlah dokumen valid berkurang menjadi 3349 dokumen. Pengurangan ini terjadi karena terdapat 206 dokumen yang tidak

memiliki jumlah token TI yang mencukupi untuk membentuk corpus LDA. Dokumen tersebut tidak digunakan karena jumlah istilah teknis yang muncul terlalu sedikit, sehingga berpotensi menghasilkan representasi topik yang lemah. Setelah pembentukan dictionary dan corpus, tidak terdapat dokumen yang gugur karena corpus kosong, sehingga seluruh 3349 dokumen tetap digunakan sebagai data final dalam pemodelan LDA.

Hasil ini menunjukkan bahwa data yang digunakan dalam model LDA tidak dipilih secara acak, melainkan melalui proses seleksi yang mempertimbangkan keterkaitan dokumen dengan domain Teknik Informatika. Dengan demikian, corpus yang digunakan dalam penelitian ini lebih relevan terhadap tujuan penelitian, yaitu memetakan tren topik skripsi mahasiswa Teknik Informatika. Penyajian data gugur juga memberikan transparansi terhadap proses seleksi data, karena setiap dokumen yang tidak digunakan memiliki alasan yang jelas dan dapat dijelaskan secara metodologis.

Tabel 4. 7 Rincian Alasan Data Gugur

Tahap Gugur	Alasan Gugur	Jumlah Dokumen
Filter awal domain TI	Tidak ditemukan sinyal istilah Teknik Informatika	406
Seleksi token TI untuk corpus LDA	Jumlah token istilah TI kurang dari batas minimal 2	201
Seleksi token TI untuk corpus LDA	Tidak memiliki token istilah TI setelah seleksi token khusus LDA	5

Berdasarkan Tabel 4.7, data yang gugur pada tahap filter awal domain TI berjumlah 406 dokumen. Dokumen tersebut gugur karena tidak ditemukan sinyal istilah Teknik Informatika yang cukup pada teks judul, kata kunci, maupun abstrak. Kondisi ini menunjukkan bahwa meskipun data diperoleh dari *repository e-theses*,

tidak seluruh dokumen memiliki representasi istilah yang sesuai dengan domain penelitian yang dianalisis.

Pada tahap seleksi token TI untuk *corpus* LDA, terdapat 201 dokumen yang gugur karena jumlah token istilah TI kurang dari batas minimal, yaitu dua token. Selain itu, terdapat 5 dokumen yang tidak memiliki token istilah TI setelah proses seleksi token khusus LDA. Dokumen-dokumen tersebut tidak digunakan karena tidak memiliki representasi istilah teknis yang cukup untuk diproses dalam pemodelan topik. Dengan demikian, penghapusan dokumen dilakukan berdasarkan kriteria yang jelas, yaitu keberadaan dan jumlah token TI yang relevan terhadap domain penelitian.

4.2 Text Processing

Tahap *Text Processing* dilakukan untuk membersihkan dan menstandarisasi data teks sebelum dilakukan pemodelan topik. Tahap ini penting karena data abstrak yang diperoleh masih mengandung banyak unsur yang tidak relevan, seperti tanda baca, angka, kata umum, dan variasi bentuk kata yang dapat mengganggu kualitas hasil *topic mining*. Oleh karena itu, diperlukan serangkaian tahapan *preprocessing* agar data menjadi lebih terstruktur dan siap digunakan oleh model *Latent Dirichlet Allocation* (LDA). ahapan *preprocessing* dalam penelitian ini meliputi *Case Folding*, *Cleaning*, *Tokenizing*, *Stopword Removal*, dan *Stemming*. Untuk memperjelas alur proses yang dilakukan, tahapan implementasi disajikan dalam bentuk pseudocode berdasarkan kode yang digunakan.

```
def preprocess_text(text):
    text = text.lower()
    text = re.sub(r'^a-zA-Z\s|', ' ', text)
    tokens = word_tokenize(text)
```

```
tokens = [word for word in tokens if word not in stopwords]
tokens = [stemmer.stem(word) for word in tokens]
return tokens
```

Proses pertama adalah *Case Folding*, yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil menggunakan fungsi `lower()`. Hal ini bertujuan untuk menghindari perbedaan penulisan kata yang sama akibat kapitalisasi. Tahap berikutnya adalah *Cleaning*, yaitu menghapus karakter yang tidak diperlukan seperti angka dan tanda baca menggunakan fungsi `regex`. Proses ini penting untuk menghilangkan noise dalam teks. Selanjutnya dilakukan *Tokenizing*, yaitu memecah teks menjadi kumpulan kata menggunakan fungsi `word_tokenize()`. Proses ini memungkinkan teks dianalisis dalam bentuk unit kata. Tahap berikutnya adalah *Stopword Removal*, yaitu menghapus kata-kata umum yang tidak memiliki makna penting, seperti “dan”, “yang”, “di”. Hal ini bertujuan untuk meningkatkan kualitas analisis topik. Terakhir dilakukan *Stemming*, yaitu mengubah kata menjadi bentuk dasarnya menggunakan `stemmer`. Proses ini bertujuan untuk menyatukan variasi kata yang memiliki makna sama.

4.2.1 Case Folding dan Cleaning

Pada tahap ini, seluruh huruf diubah menjadi huruf kecil (lowercase) agar kata yang sama tidak dianggap berbeda hanya karena perbedaan kapitalisasi. Selain itu, dilakukan penghapusan angka, tanda baca, URL, dan karakter khusus.

Tabel 4. 8 *Case Folding dan Cleaning*

No	Sebelum	Sesudah
1	Analisis Sistem Informasi Berbasis Web Tahun 2024!!!	Analisis Sistem Informasi Berbasis Web Tahun
2	Implementasi Algoritma Genetika Pada Data	Implementasi Algoritma Genetika Pada Data

No	Sebelum	Sesudah
3	Pengujian Sistem 100% Berhasil	Pengujian Sistem Berhasil
4	Model Cnn Untuk Klasifikasi Citra!	Model Cnn Untuk Klasifikasi Citra
5	Sistem Informasi Berbasis Android V2.0	Sistem Informasi Berbasis Android V
6	Deteksi Objek Menggunakan Yolov5	Deteksi Objek Menggunakan Yolov
7	Analisis Data Covid-19 Tahun 2021	Analisis Data Covid Tahun
8	Sistem Pakar Diagnosa Penyakit!!!	Sistem Pakar Diagnosa Penyakit
9	Aplikasi Web Dengan Framework Laravel	Aplikasi Web Dengan Framework Laravel
10	Optimasi Algoritma Pso Pada Data Besar	Optimasi Algoritma Pso Pada Data Besar

Tabel 4.8 menunjukkan bahwa proses *Case Folding* berhasil mengubah seluruh huruf menjadi huruf kecil, sedangkan *Cleaning* menghapus angka, simbol, dan tanda baca. Hal ini membuat teks menjadi lebih bersih dan konsisten, sehingga mengurangi noise sebelum masuk tahap berikutnya.

4.2.2 Tokenizing

Setelah teks dibersihkan, langkah berikutnya adalah memecah kalimat menjadi kumpulan kata atau token.

Tabel 4. 9 *Tokenizing*

No	Sebelum	Sesudah
1	analisis sistem informasi berbasis web	[analisis, sistem, informasi, berbasis, web]
2	implementasi algoritma genetika pada data	[implementasi, algoritma, genetika, pada, data]
3	pengujian sistem berhasil	[pengujian, sistem, berhasil]
4	model cnn untuk klasifikasi citra	[model, cnn, untuk, klasifikasi, citra]
5	sistem informasi berbasis android	[sistem, informasi, berbasis, android]
6	deteksi objek menggunakan yolov	[deteksi, objek, menggunakan, yolov]
7	analisis data covid	[analisis, data, covid]
8	sistem pakar diagnosa penyakit	[sistem, pakar, diagnosa, penyakit]
9	aplikasi web framework laravel	[aplikasi, web, framework, laravel]
10	optimasi algoritma pso data besar	[optimasi, algoritma, pso, data, besar]

Tabel 4.9 menunjukkan proses *Tokenizing*, di mana kalimat dipecah menjadi kata-kata individual. Hal ini memungkinkan model untuk menganalisis teks pada tingkat kata sehingga memudahkan proses pembentukan topik.

4.2.3 Stopword Removal

Tahap berikutnya adalah menghapus stopwords, yaitu kata-kata umum yang sering muncul tetapi tidak memberikan kontribusi besar terhadap pembentukan topik.

Tabel 4. 10 *Stopword Removal*

No	Sebelum	Sesudah
1	[analisis, sistem, informasi, berbasis, web]	[analisis, sistem, informasi, berbasis, web]
2	[implementasi, algoritma, genetika, pada, data]	[implementasi, algoritma, genetika, data]
3	[model, cnn, untuk, klasifikasi, citra]	[model, cnn, klasifikasi, citra]
4	[deteksi, objek, menggunakan, yolov]	[deteksi, objek, yolov]
5	[analisis, data, covid]	[analisis, covid]
6	[sistem, pakar, diagnosa, penyakit]	[sistem, pakar, diagnosa, penyakit]
7	[aplikasi, web, framework, laravel]	[aplikasi, web, framework, laravel]
8	[optimasi, algoritma, pso, data, besar]	[optimasi, algoritma, pso, besar]
9	[pengujian, sistem, berhasil]	[pengujian, sistem, berhasil]
10	[analisis, sistem, yang, digunakan]	[analisis, sistem, digunakan]

Tabel 4.10 menunjukkan bahwa kata-kata umum seperti “pada”, “untuk”, dan “yang” berhasil dihapus. Penghapusan stopwords membantu model fokus pada kata-kata penting yang berkontribusi terhadap pembentukan topik.

4.2.4 Stemming

Stemming dilakukan untuk mengubah kata menjadi bentuk dasarnya. Sebagai contoh seperti pada tabel 4.11 berikut.

Tabel 4. 11 *Stemming*

No	Sebelum	Sesudah
1	analisis, sistem, informasi	analisis, sistem, informasi
2	implementasi, algoritma, genetika	implementasi, algoritma, genetika

No	Sebelum	Sesudah
3	menggunakan, dilakukan, penelitian	guna, laku, teliti
4	pengujian, sistem, berhasil	uji, sistem, hasil
5	deteksi, objek, menggunakan	deteksi, objek, guna
6	klasifikasi, citra, digital	klasifikasi, citra, digital
7	optimasi, algoritma, berjalan	optimasi, algoritma, jalan
8	pengolahan, data, besar	olah, data, besar
9	perancangan, sistem, informasi	rancang, sistem, informasi
10	pengembangan, aplikasi, web	kembang, aplikasi, web

Tabel 4.11 menunjukkan bahwa kata dengan imbuhan berhasil diubah menjadi bentuk dasar. Proses ini mengurangi variasi kata sehingga model dapat lebih mudah mengenali hubungan antar kata dalam pembentukan topik.

Berdasarkan hasil *preprocessing*, data yang awalnya tidak terstruktur berhasil dibersihkan menjadi bentuk yang lebih konsisten. Setiap tahap memberikan kontribusi penting, di mana *Cleaning* menghilangkan noise, *Tokenizing* memecah teks, *Stopword Removal* menghapus kata tidak penting, dan *Stemming* menyederhanakan kata. Kombinasi dari tahapan ini menghasilkan data yang lebih optimal untuk digunakan dalam pemodelan LDA.

4.3 Membangun *Corpus*

Setelah tahap *preprocessing* selesai dilakukan pada data teks, langkah selanjutnya adalah membangun *Corpus* yang akan digunakan sebagai input pada model *Latent Dirichlet Allocation* (LDA). Pada tahap ini, data teks yang telah dibersihkan dan dinormalisasi tidak lagi digunakan dalam bentuk teks mentah, melainkan diubah menjadi representasi numerik agar dapat diproses oleh algoritma pembelajaran mesin.

Proses pembangunan *Corpus* dalam penelitian ini terdiri dari beberapa tahapan utama, yaitu pembentukan bigram, pembentukan *Dictionary*, *filtering* kata,

serta pembentukan *Corpus* dalam bentuk *bag-of-words*. Setiap tahapan memiliki peran penting dalam meningkatkan kualitas representasi data sehingga model LDA dapat menghasilkan topik yang lebih akurat dan interpretatif.

Corpus pada penelitian ini dibentuk dari 3349 dokumen yang telah lolos tahap filter awal domain Teknik Informatika dan seleksi token TI. Dokumen yang digunakan dalam corpus merupakan dokumen yang memiliki token istilah TI yang cukup serta tidak menghasilkan corpus kosong setelah proses pembentukan dictionary. Dengan demikian, corpus yang digunakan tidak hanya merepresentasikan kumpulan teks hasil preprocessing, tetapi juga telah disaring agar sesuai dengan domain penelitian. Proses ini dilakukan untuk memastikan bahwa model LDA mempelajari pola kata yang relevan dengan bidang Teknik Informatika, sehingga hasil topik yang diperoleh lebih sesuai dengan tujuan penelitian.

4.3.1 Bigram

Pada tahap ini dilakukan pembentukan bigram, yaitu penggabungan dua kata yang sering muncul secara berurutan dalam dokumen. Tujuan dari penggunaan bigram adalah untuk mempertahankan konteks kata sehingga makna yang dihasilkan menjadi lebih spesifik.

```
bigram = Phrases(df_ti["processed_tokens"], min_count=3,
threshold=8)
bigram_phraser = Phraser(bigram)
df_ti["processed_bigram"] =
df_ti["processed_tokens"].apply(lambda x: bigram_phraser[x])
processed_texts = df_ti["processed_bigram"].tolist()
```

Proses bigram dilakukan untuk menggabungkan kata-kata yang sering muncul secara berurutan sehingga membentuk satu kesatuan makna. Parameter

`min_count=3` menunjukkan bahwa pasangan kata harus muncul minimal tiga kali untuk dianggap sebagai bigram, sedangkan `threshold=8` mengatur tingkat sensitivitas pembentukan bigram. Hasil dari proses ini adalah kumpulan token baru yang telah mempertimbangkan hubungan antar kata.

Tabel 4. 12 Hasil *Bigram*

No	Sebelum	Sesudah
1	[web, service, android, api]	[web_service, android, api]
2	[algoritma, genetika, jadwal]	[algoritma_genetika, jadwal]
3	[klasifikasi, naive, bayes]	[klasifikasi_naive_bayes]
4	[machine, learning, model]	[machine_learning, model]
5	[sistem, informasi, web]	[sistem_informasi, web]

Tabel 4.12 menunjukkan bahwa pasangan kata yang sering muncul bersama berhasil digabungkan menjadi satu token baru. Sebagai contoh, kata *web* dan *service* digabung menjadi *web_service*, serta *machine* dan *learning* menjadi *machine_learning*. Proses ini bertujuan untuk mempertahankan makna kontekstual dari dua kata yang saling berkaitan sehingga tidak terpisah dalam analisis.

Selain itu, penggunaan bigram memberikan dampak yang signifikan terhadap kualitas representasi teks. Tanpa *bigram*, model LDA akan menganggap setiap kata berdiri sendiri sehingga hubungan antar kata dapat hilang. Dengan adanya bigram, istilah-istilah penting yang sering muncul dalam domain tertentu, seperti teknologi informasi, dapat dipertahankan sebagai satu kesatuan konsep yang utuh.

Dari sisi analisis, penerapan bigram membantu meningkatkan interpretasi topik yang dihasilkan. Topik yang terbentuk menjadi lebih spesifik dan mudah dipahami karena mengandung istilah gabungan yang memiliki makna lebih jelas dibandingkan kata tunggal. Oleh karena itu, penggunaan bigram dalam penelitian

ini menjadi salah satu langkah penting dalam meningkatkan kualitas hasil topic modeling.

4.3.2 Dictionary

Dictionary merupakan struktur data yang digunakan untuk memetakan setiap kata unik ke dalam ID numerik. *Dictionary* berfungsi untuk mengubah kata menjadi representasi numerik. Setiap kata unik diberikan ID sehingga dapat digunakan dalam proses pembentukan *Corpus*.

Tabel 4. 13 Contoh *Dictionary*

No	Kata	ID
1	algoritma	0
2	klasifikasi	1
3	sistem	2
4	jaringan	3
5	data	4
6	citra	5
7	web_service	6
8	machine_learning	7
9	aplikasi	8
10	model	9

Tabel 4.13 menunjukkan bahwa setiap kata unik yang terdapat dalam dataset diberikan sebuah ID numerik yang bersifat unik. Proses ini memungkinkan teks yang awalnya tidak terstruktur dapat diubah menjadi bentuk numerik yang dapat diproses oleh model LDA. Setiap ID merepresentasikan satu kata tertentu dalam keseluruhan *Corpus*.

Selain itu, *Dictionary* berperan sebagai dasar dalam pembentukan *Corpus*, karena seluruh kata dalam dokumen akan direferensikan berdasarkan ID yang telah ditentukan. Hal ini membuat proses perhitungan frekuensi kata menjadi lebih

efisien dan terstruktur. Dengan adanya *Dictionary*, model tidak perlu membaca kata secara langsung, tetapi cukup menggunakan ID sebagai representasi.

Dari sudut pandang analisis, ukuran *Dictionary* yang terbentuk juga mencerminkan kompleksitas data. Semakin banyak kata unik yang tersisa setelah proses filtering, maka semakin beragam informasi yang dapat ditangkap oleh model. Namun demikian, jumlah kata juga harus tetap dikontrol agar tidak mengandung terlalu banyak noise yang dapat menurunkan kualitas topik.

4.3.3 Filtering Kata

Filtering dilakukan untuk meningkatkan kualitas data dengan menghapus kata yang tidak memberikan kontribusi terhadap pembentukan topik.

Tabel 4. 14 Contoh *Filtering* Kata

No	Kata	Status
1	algoritma	dipertahankan
2	klasifikasi	dipertahankan
3	sistem	dipertahankan
4	hasil	dihapus
5	nilai	dihapus
6	baik	dihapus
7	uji	dihapus
8	data	dipertahankan
9	model	dipertahankan
10	metode	dihapus

Tabel 4.14 menunjukkan hasil proses *filtering* kata, di mana kata-kata tertentu dipertahankan atau dihapus berdasarkan kriteria yang telah ditentukan. Kata seperti *algoritma*, *klasifikasi*, dan *sistem* dipertahankan karena memiliki makna penting dalam konteks penelitian, sedangkan kata seperti *hasil*, *nilai*, dan *baik* dihapus karena bersifat umum dan tidak spesifik.

Proses *filtering* dilakukan dengan dua pendekatan, yaitu otomatis dan manual. *Filtering* otomatis menggunakan parameter frekuensi kemunculan kata, sedangkan *filtering* manual dilakukan berdasarkan pengetahuan domain untuk menghapus kata-kata yang tidak relevan. Kombinasi kedua pendekatan ini bertujuan untuk menghasilkan kumpulan kata yang lebih berkualitas.

Secara analitis, proses *filtering* sangat berpengaruh terhadap kualitas topik yang dihasilkan oleh model LDA. Jika terlalu banyak kata umum yang dipertahankan, maka topik yang terbentuk akan menjadi kurang jelas dan cenderung ambigu. Sebaliknya, dengan menghilangkan kata-kata yang tidak relevan, model dapat lebih fokus pada kata-kata penting yang benar-benar merepresentasikan isi dokumen.

4.3.4 Corpus

Corpus digunakan untuk merepresentasikan dokumen dalam bentuk numerik, yaitu pasangan antara ID kata dan jumlah kemunculan.

Tabel 4. 15 Contoh *Corpus*

Dokumen	Representasi
D1	[(0,1), (1,3), (2,2), (3,3), (4,8)]
D2	[(6,5), (7,1)]
D3	[(0,1), (8,4), (9,4)]
D4	[(0,1), (12,1), (13,2)]
D5	[(8,1), (17,1), (18,4)]

Tabel 4.15 menunjukkan representasi dokumen dalam bentuk *Bag of Words* (BoW), yaitu pasangan antara ID kata dan jumlah kemunculannya dalam dokumen. Sebagai contoh, pasangan (0,1) menunjukkan bahwa kata dengan ID 0 muncul sebanyak satu kali dalam dokumen tersebut, sedangkan (1,3) menunjukkan kemunculan sebanyak tiga kali.

Representasi ini sangat penting karena model LDA bekerja berdasarkan distribusi frekuensi kata dalam dokumen. Dengan menggunakan format *Bag of Words*, setiap dokumen dapat direpresentasikan secara numerik tanpa memperhatikan urutan kata, sehingga memudahkan proses perhitungan probabilitas dalam model.

Dari sisi analisis, *Corpus* yang terbentuk mencerminkan struktur data yang siap digunakan dalam pemodelan topik. Setiap dokumen telah direpresentasikan secara sistematis sehingga model dapat mengidentifikasi pola kemunculan kata dan mengelompokkan dokumen ke dalam topik tertentu. Oleh karena itu, tahap pembentukan *Corpus* menjadi langkah krusial sebelum masuk ke proses pemodelan LDA.

4.4 Pemodelan LDA

Pada tahap ini dilakukan proses pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA). Metode ini digunakan untuk mengidentifikasi topik-topik tersembunyi dalam kumpulan dokumen skripsi mahasiswa berdasarkan distribusi kemunculan kata.

LDA merupakan model probabilistik generatif yang mengasumsikan bahwa setiap dokumen terdiri dari kombinasi beberapa topik, dan setiap topik terdiri dari distribusi kata tertentu. Dengan pendekatan ini, model dapat mengelompokkan dokumen berdasarkan kemiripan kata sehingga terbentuk topik-topik yang merepresentasikan tema penelitian.

4.4.1 Proses Pemodelan LDA

Pemodelan dilakukan menggunakan library Gensim dengan memanfaatkan *Corpus* dan *Dictionary* yang telah dibentuk pada tahap sebelumnya.

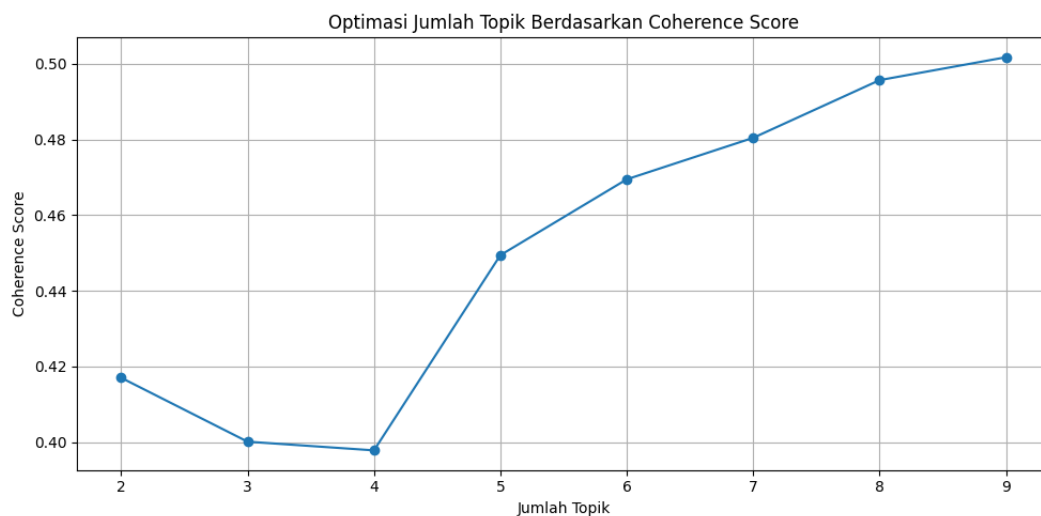
```
lda_model = LdaModel(
    Corpus=Corpus,
    id2word=Dictionary,
    num_topics=9,
    passes=10,
    random_state=42
)
```

Pseudocode di atas menunjukkan proses pembentukan model LDA dengan parameter utama jumlah topik sebanyak 9. Parameter *Corpus* merupakan representasi dokumen dalam bentuk *Bag of Words*, sedangkan *Dictionary* berfungsi sebagai pemetaan kata ke ID. Parameter *passes* digunakan untuk menentukan jumlah iterasi pelatihan model agar distribusi topik menjadi lebih stabil, dan *random_state* digunakan untuk memastikan hasil yang konsisten.

Pemilihan jumlah topik dilakukan berdasarkan proses optimasi menggunakan *Coherence Score*, sehingga model yang dihasilkan tidak hanya berdasarkan asumsi, tetapi melalui proses evaluasi yang sistematis.

4.4.2 Optimasi Jumlah Topik

Penentuan jumlah topik pada model *Latent Dirichlet Allocation* (LDA) dilakukan melalui pengujian beberapa variasi jumlah topik, yaitu mulai dari $K = 2$ sampai $K = 9$. Setiap model yang dihasilkan kemudian dievaluasi menggunakan *coherence score* untuk mengukur tingkat keterkaitan semantik antar kata dalam setiap topik. Semakin tinggi nilai *coherence*, maka semakin baik kualitas topik yang dihasilkan, karena kata-kata dalam topik tersebut dianggap semakin konsisten dan mudah diinterpretasikan.



Gambar 4. 1 Optimasi Jumlah Topik

Berdasarkan Gambar 4.1, terlihat bahwa nilai *coherence score* cenderung mengalami peningkatan seiring dengan bertambahnya jumlah topik, meskipun peningkatannya tidak berlangsung secara sepenuhnya linier. Pada saat jumlah topik $K = 2$, nilai *coherence* masih relatif rendah, yaitu sebesar 0,290132. Nilai ini kemudian meningkat pada $K = 3$ menjadi 0,344412, dan kembali meningkat pada $K = 4$ menjadi 0,372782. Kenaikan tersebut menunjukkan bahwa ketika jumlah topik ditambah, model mulai mampu memisahkan dokumen ke dalam kelompok tema yang lebih spesifik dibandingkan jumlah topik yang terlalu sedikit.

Peningkatan nilai *coherence* masih berlanjut pada $K = 5$, dengan skor sebesar 0,397428. Namun, pada $K = 6$ terjadi sedikit penurunan menjadi 0,377348, yang menunjukkan bahwa penambahan jumlah topik tidak selalu menghasilkan kualitas topik yang lebih baik. Setelah itu, nilai *coherence* kembali meningkat pada $K = 7$ sebesar 0,385226, dan mengalami kenaikan yang cukup signifikan pada $K = 8$, yaitu mencapai 0,443867. Nilai ini merupakan *coherence score* tertinggi dari seluruh skenario pengujian yang dilakukan.

Meskipun nilai *coherence* tertinggi diperoleh pada $K = 8$, penelitian ini menetapkan $K = 9$ sebagai model final, sebagaimana ditunjukkan oleh garis vertikal putus-putus pada grafik. Keputusan ini diambil karena nilai *coherence* pada $K = 9$ sebesar 0,429723 masih tergolong tinggi dan memiliki selisih yang relatif kecil terhadap nilai terbaik pada $K = 8$, yaitu sekitar 0,014144. Dengan selisih yang tidak terlalu besar tersebut, penggunaan 9 topik masih dinilai layak karena mampu memberikan pemisahan tema yang lebih rinci dan lebih sesuai dengan kebutuhan interpretasi topik dalam penelitian ini.

Secara substantif, pemilihan $K = 9$ juga memberikan keuntungan dalam proses analisis lanjutan, karena hasil topik yang terbentuk menjadi lebih detail dan lebih representatif untuk menggambarkan keragaman tema penelitian skripsi mahasiswa Teknik Informatika. Dengan demikian, keputusan penggunaan 9 topik tidak hanya didasarkan pada nilai evaluasi kuantitatif, tetapi juga mempertimbangkan aspek interpretabilitas, keterbacaan hasil, serta kesesuaian dengan kebutuhan analisis pada penelitian ini.

Berdasarkan hasil evaluasi tersebut, dapat disimpulkan bahwa proses optimasi jumlah topik menunjukkan model LDA telah mampu membentuk topik-topik yang cukup koheren. Nilai *coherence* yang berada pada rentang 0,29 hingga 0,44 menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengelompokkan kata-kata ke dalam topik yang bermakna. Oleh karena itu, model final dengan 9 topik digunakan pada tahap analisis berikutnya, yaitu interpretasi topik, visualisasi distribusi topik menggunakan pyLDAvis, serta analisis tren topik penelitian mahasiswa.

4.4.3 Hasil Topik LDA

Setelah model terbentuk, dilakukan ekstraksi topik untuk mengetahui kata-kata dominan pada setiap topik.

Tabel 4. 16 Hasil Topik LDA

Topik	Kata Dominan
Topik 1	algoritma, enkripsi, kriptografi, ahp, saw
Topik 2	network, citra, image, cnn, deep learning
Topik 3	graf, svm, random forest, machine learning
Topik 4	fuzzy, prediksi, objek
Topik 5	java, decision, topsi
Topik 6	klasifikasi, naive bayes, text mining
Topik 7	deteksi, optimasi, dijkstra
Topik 8	clustering, cluster, learning
Topik 9	sensor, iot, arduino, monitoring

Tabel 4.16 atas menunjukkan hasil ekstraksi topik dari model LDA, di mana setiap topik direpresentasikan oleh kata-kata dominan dengan probabilitas tinggi. Kata-kata tersebut mencerminkan tema utama dari masing-masing topik dalam dataset skripsi mahasiswa. Secara umum, topik yang dihasilkan mencakup berbagai bidang dalam Teknik Informatika, seperti keamanan data, pengolahan citra, machine learning, sistem pendukung keputusan, clustering, hingga Internet of Things. Selain itu, kata-kata teknis seperti *cnn*, *naive bayes*, *iot*, dan *clustering* menunjukkan bahwa model mampu mengenali istilah-istilah penting yang sering digunakan dalam penelitian, sehingga topik yang dihasilkan menjadi lebih relevan dan mudah diinterpretasikan.

4.5 Evaluasi Model

Pada tahap ini dilakukan evaluasi terhadap model *Latent Dirichlet Allocation* (LDA) yang telah dibangun. Evaluasi bertujuan untuk mengukur

kualitas topik yang dihasilkan serta memastikan bahwa model mampu merepresentasikan data secara optimal. Dalam penelitian ini, evaluasi dilakukan menggunakan metrik *Coherence Score* serta analisis distribusi topik berdasarkan hasil visualisasi.

Coherence Score digunakan untuk mengukur tingkat keterkaitan antar kata dalam suatu topik. Semakin tinggi nilai *coherence*, maka semakin baik kualitas topik yang dihasilkan. Selain itu, evaluasi juga dilakukan dengan melihat distribusi topik untuk mengetahui keseimbangan representasi dokumen dalam setiap topik.

4.5.1 Evaluasi Berdasarkan *Coherence Score*

Evaluasi pertama dilakukan menggunakan *Coherence Score* yang diperoleh dari hasil pengujian berbagai jumlah topik.

Tabel 4. 17 Nilai *Coherence* Berdasarkan Jumlah Topik

Jumlah Topik	<i>Coherence Score</i>
2	0.290132
3	0.344412
4	0.372782
5	0.397428
6	0.377348
7	0.385226
8	0.443867
9	0.429723

Tabel 4.17 atas menunjukkan bahwa nilai *coherence* cenderung meningkat seiring dengan bertambahnya jumlah topik. Pada jumlah topik yang kecil seperti pada topik 2, nilai *coherence* relatif rendah karena model belum mampu memisahkan dokumen ke dalam kelompok topik yang lebih spesifik. Hal ini menyebabkan beberapa tema penelitian masih tercampur dalam satu topik yang sama.

Seiring bertambahnya jumlah topik, model menjadi lebih mampu menangkap variasi tema yang terdapat dalam dataset. Hal ini terlihat dari peningkatan nilai *coherence* pada jumlah topik 5 hingga 9. Peningkatan ini menunjukkan bahwa model semakin baik dalam mengelompokkan kata-kata yang memiliki keterkaitan semantik dalam satu topik.

Nilai *coherence* tertinggi diperoleh pada jumlah topik 8, yaitu sebesar 0.443867. Hal ini menunjukkan bahwa konfigurasi tersebut merupakan model terbaik dalam penelitian ini, karena mampu menghasilkan topik yang paling konsisten dan mudah diinterpretasikan.

4.5.2 Evaluasi Berdasarkan Distribusi Topik

Selain menggunakan *Coherence Score*, evaluasi juga dilakukan dengan melihat distribusi topik berdasarkan hasil visualisasi pyLDAvis.

Tabel 4. 18 Distribusi Proporsi Topik

Topik	Proporsi
Topik 1	17.46%
Topik 2	16.36%
Topik 3	12.99%
Topik 4	12.88%
Topik 5	±10%
Topik 6	±9%
Topik 7	±8%
Topik 8	±7%
Topik 9	4.01%

Tabel 4.18 menunjukkan bahwa distribusi topik dalam *Corpus* tidak merata, di mana beberapa topik memiliki proporsi yang lebih besar dibandingkan yang lain. Topik 1 dan Topik 2 merupakan topik yang paling dominan, sedangkan Topik 9 memiliki proporsi paling kecil.

Distribusi ini menunjukkan bahwa terdapat beberapa tema penelitian yang lebih sering muncul dalam dataset skripsi mahasiswa, seperti algoritma, klasifikasi, dan pengolahan data. Sementara itu, topik dengan proporsi kecil seperti Internet of Things (IoT) menunjukkan bahwa tema tersebut masih relatif jarang dibandingkan tema lainnya.

Meskipun distribusi tidak merata, kondisi ini masih dapat diterima karena mencerminkan kondisi nyata dari data penelitian. Model LDA tidak memaksakan distribusi yang sama, melainkan mengikuti pola kemunculan data yang sebenarnya.

4.5.3 Analisis Kualitas Model

Berdasarkan hasil evaluasi yang telah dilakukan, model LDA yang dibangun dalam penelitian ini dapat dikatakan memiliki kualitas yang cukup baik. Hal ini ditunjukkan oleh nilai *coherence* yang relatif tinggi serta distribusi topik yang mampu merepresentasikan variasi tema penelitian dalam dataset.

Penggunaan *preprocessing* yang ketat, pembentukan *bigram*, serta *filtering* kata terbukti memberikan kontribusi terhadap peningkatan kualitas model. Dengan adanya tahapan tersebut, kata-kata yang tidak relevan dapat diminimalkan sehingga model dapat lebih fokus pada kata-kata yang memiliki makna penting.

Selain itu, visualisasi menggunakan pyLDavis menunjukkan bahwa sebagian besar topik telah terpisah dengan cukup baik, meskipun masih terdapat beberapa topik yang berdekatan. Hal ini menunjukkan bahwa model telah mampu mengelompokkan dokumen berdasarkan tema, namun masih terdapat keterkaitan antar beberapa topik.

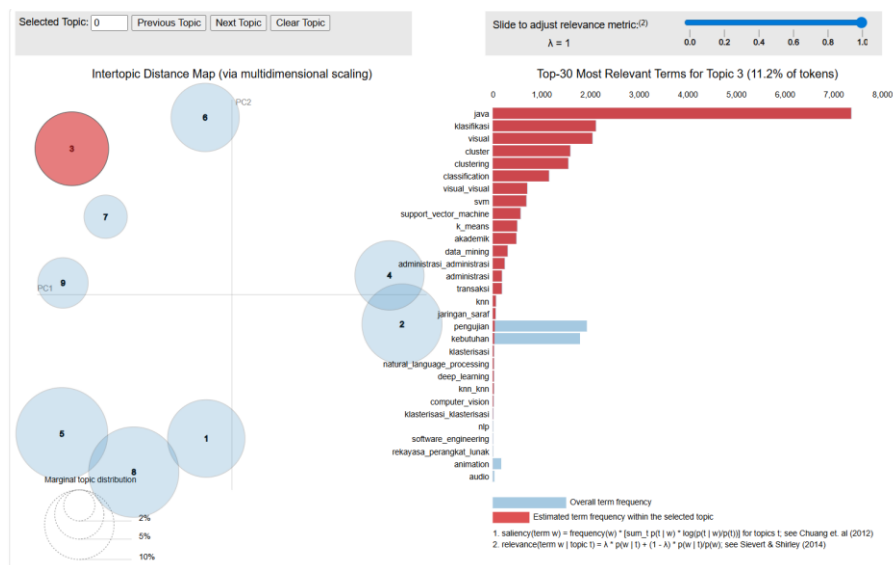
Dengan demikian, model yang telah dibangun dapat digunakan sebagai dasar dalam proses analisis lebih lanjut pada tahap pembahasan, khususnya dalam mengidentifikasi tren penelitian dan karakteristik topik yang dihasilkan.

4.6 Hasil *Topic mining*

Pada tahap ini dilakukan visualisasi terhadap hasil pemodelan topik menggunakan metode LDA. Visualisasi bertujuan untuk memberikan gambaran yang lebih intuitif mengenai distribusi topik, hubungan antar topik, serta dominasi tema penelitian dalam dataset.

4.6.1 Visualisasi Distribusi Topik (pyLDavis)

Visualisasi distribusi topik dilakukan menggunakan *library* pyLDavis untuk menggambarkan hubungan antar topik serta proporsi masing-masing topik dalam *Corpus*. Visualisasi ini menyajikan topik dalam bentuk lingkaran pada bidang dua dimensi berdasarkan hasil reduksi dimensi, di mana setiap lingkaran merepresentasikan satu topik, seperti pada gambar 4.1 berikut.



Gambar 4. 2 Visualisasi *pyLDavis*

Gambar 4.2 menunjukkan bahwa model LDA menghasilkan sembilan topik yang tersebar pada bidang dua dimensi. Pada panel sebelah kiri, setiap lingkaran menunjukkan satu topik yang terbentuk dari hasil pemodelan LDA. Posisi lingkaran menggambarkan kedekatan atau perbedaan distribusi kata antar topik. Topik yang letaknya saling berdekatan menunjukkan bahwa terdapat kemiripan kata atau konteks pembahasan antar topik, sedangkan topik yang berjauhan menunjukkan bahwa topik tersebut memiliki karakteristik kata yang lebih berbeda.

Berdasarkan visualisasi, sebagian besar topik telah tersebar pada area yang berbeda, meskipun masih terdapat beberapa topik yang saling berdekatan. Misalnya, Topik 2 dan Topik 4 terlihat berdekatan, sehingga dapat diinterpretasikan bahwa kedua topik tersebut memiliki kemiripan distribusi kata. Kondisi ini wajar terjadi dalam pemodelan topik, karena beberapa bidang dalam Teknik Informatika sering memiliki irisan istilah, misalnya antara sistem informasi, klasifikasi, data mining, pengujian, dan kebutuhan sistem. Sebaliknya, topik yang berada lebih jauh menunjukkan bahwa kelompok kata pada topik tersebut memiliki perbedaan konteks yang lebih kuat.

Pada gambar tersebut, Topik 3 sedang dipilih dan memiliki proporsi sekitar 11,2% dari keseluruhan token. Panel sebelah kanan menampilkan 30 kata paling relevan pada Topik 3. Kata-kata yang muncul antara lain *java*, *klasifikasi*, *visual*, *cluster*, *clustering*, *classification*, *svm*, *support_vector_machine*, *k_means*, *data_mining*, *knn*, *jaringan_saraf*, *natural_language_processing*, *deep_learning*, *nlp*, *software_engineering*, dan *rekayasa_perangkat_lunak*. Kemunculan istilah-istilah tersebut menunjukkan bahwa Topik 3 memiliki keterkaitan kuat dengan

bidang *Intelligence System*, khususnya pada tema klasifikasi, *clustering*, *data mining*, dan *machine learning*.

Pada panel kanan, warna merah menunjukkan estimasi frekuensi kata pada topik yang sedang dipilih, sedangkan warna biru menunjukkan frekuensi kata secara keseluruhan dalam *corpus*. Apabila batang merah suatu kata lebih dominan, maka kata tersebut lebih khas terhadap topik yang sedang dipilih. Sebaliknya, apabila batang biru lebih besar, maka kata tersebut sering muncul dalam *corpus* secara umum, tetapi belum tentu menjadi ciri khusus topik tersebut. Dalam visualisasi ini, kata seperti *java* memiliki frekuensi yang sangat tinggi pada Topik 3, sedangkan kata seperti *penguji* dan *kebutuhan tampak* memiliki frekuensi umum yang cukup besar, tetapi tidak menjadi kata paling dominan pada topik terpilih.

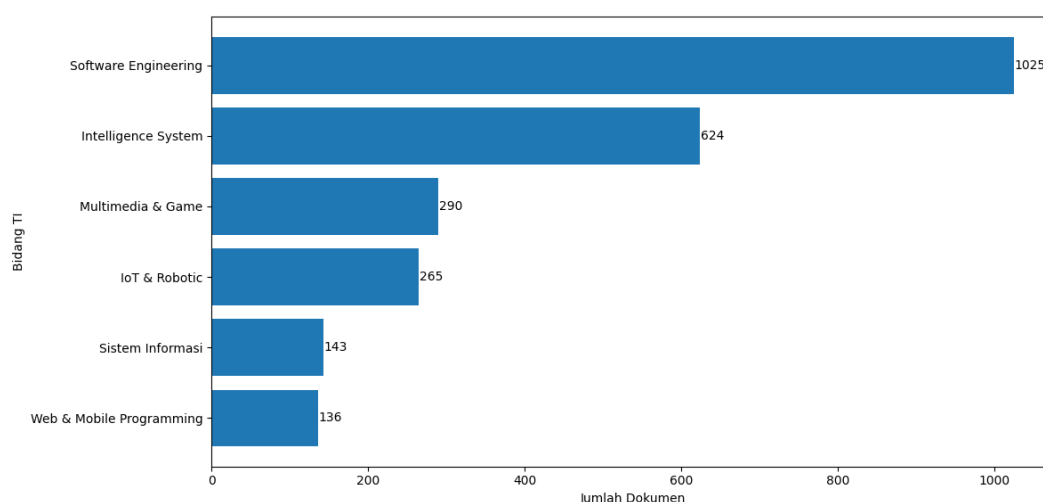
Pada Gambar 4.2, nilai $\lambda = 1$, sehingga kata yang ditampilkan lebih diprioritaskan berdasarkan probabilitas kemunculannya dalam topik terpilih. Artinya, kata-kata yang muncul di panel kanan merupakan kata yang paling sering dan paling dominan pada topik tersebut. Dengan demikian, visualisasi *pyLDAvis* tidak hanya menunjukkan sebaran topik, tetapi juga membantu dalam memahami karakteristik setiap topik berdasarkan kata-kata yang paling relevan.

Secara keseluruhan, visualisasi *pyLDAvis* menunjukkan bahwa model LDA mampu membentuk struktur topik yang dapat diinterpretasikan. Beberapa topik memiliki posisi yang cukup terpisah, sehingga menunjukkan adanya perbedaan tema antar kelompok dokumen. Meskipun terdapat topik yang berdekatan, hal tersebut menunjukkan adanya irisan konteks yang wajar dalam penelitian Teknik Informatika. Dengan demikian, hasil visualisasi ini dapat

digunakan sebagai dasar untuk menganalisis karakteristik topik, kata dominan, serta tren bidang penelitian mahasiswa.

4.6.2 Visualisasi Tren Topik Penelitian

Distribusi topik penelitian dianalisis berdasarkan hasil pengelompokan dokumen ke dalam kategori bidang Teknik Informatika yang telah disesuaikan dengan struktur bidang keilmuan pada objek penelitian. Proses pengelompokan ini dilakukan dengan cara mengklasifikasikan setiap dokumen ke dalam satu bidang berdasarkan topik dominan yang diperoleh dari hasil pemodelan *Latent Dirichlet Allocation* (LDA).



Gambar 4. 3 Distribusi Topik Bidang TI

Berdasarkan Gambar 4.3, terlihat bahwa distribusi topik skripsi mahasiswa berdasarkan bidang TI tidak merata. Dari total 2090 dokumen yang berhasil dipetakan ke dalam enam bidang Teknik Informatika, bidang *Software Engineering* menempati posisi tertinggi dengan jumlah 583 dokumen atau 27,89%. Jumlah ini menunjukkan bahwa penelitian mahasiswa cukup banyak berfokus pada aspek rekayasa perangkat lunak, seperti perancangan sistem, pengujian perangkat lunak,

pengembangan aplikasi, kebutuhan sistem, model pengembangan, serta evaluasi kualitas perangkat lunak.

Bidang kedua yang memiliki jumlah hampir sama adalah *Intelligence System* dengan 579 dokumen atau 27,70%. Bidang ini mencakup penelitian yang berkaitan dengan kecerdasan buatan, *machine learning*, *data mining*, klasifikasi, prediksi, *clustering*, *natural language processing*, dan metode komputasi cerdas lainnya. Selisih antara *Software Engineering* dan *Intelligence System* hanya 4 dokumen, sehingga dapat dikatakan bahwa kedua bidang tersebut menjadi dua tema penelitian paling dominan dalam dataset. Dominasi ini menunjukkan bahwa arah penelitian mahasiswa Teknik Informatika banyak berfokus pada pengembangan sistem sekaligus penerapan metode cerdas untuk menyelesaikan permasalahan berbasis data.

Selanjutnya, bidang *IoT & Robotic* menempati posisi ketiga dengan jumlah 359 dokumen atau 17,18%. Jumlah ini menunjukkan bahwa penelitian terkait *Internet of Things*, sensor, *monitoring*, mikrokontroler, robotika, dan sistem otomatisasi juga memiliki kontribusi yang cukup besar. Bidang ini menunjukkan adanya minat mahasiswa dalam mengembangkan sistem yang menghubungkan perangkat keras dan perangkat lunak, terutama pada aplikasi berbasis sensor, kontrol, dan pemantauan.

Bidang Multimedia & Game berada pada posisi berikutnya dengan jumlah 281 dokumen atau 13,44%. Bidang ini mencakup tema seperti game, animasi, multimedia, citra digital, visualisasi, *augmented reality*, *virtual reality*, serta pengolahan media digital. Jumlah tersebut menunjukkan bahwa bidang multimedia

masih menjadi salah satu topik yang cukup diminati, meskipun tidak sebesar *Software Engineering* dan *Intelligence System*. Hal ini dapat menunjukkan bahwa penelitian multimedia cenderung lebih spesifik dan biasanya berkaitan dengan pengembangan media interaktif, visualisasi, atau aplikasi berbasis pengalaman pengguna.

Sementara itu, bidang Sistem Informasi memiliki jumlah 183 dokumen atau 8,76%. Bidang ini mencakup penelitian yang berkaitan dengan pengelolaan informasi, database, dashboard, sistem administrasi, sistem pendukung keputusan, serta aplikasi manajemen. Jumlah ini menunjukkan bahwa penelitian pada bidang Sistem Informasi tetap muncul dalam dataset, meskipun proporsinya lebih rendah dibandingkan bidang *Software Engineering*, *Intelligence System*, *IoT & Robotic*, dan *Multimedia & Game*. Hal ini dapat terjadi karena beberapa penelitian sistem informasi juga memiliki irisan dengan bidang lain, terutama *Software Engineering* dan *Web & Mobile Programming*.

Bidang dengan jumlah paling sedikit adalah *Web & Mobile Programming*, yaitu 105 dokumen atau 5,02%. Bidang ini mencakup penelitian yang berhubungan dengan pengembangan aplikasi *web*, *mobile*, *API*, *Android*, *Laravel*, *Flutter*, dan teknologi pemrograman sejenis. Proporsi yang lebih kecil tidak berarti bidang ini tidak penting, tetapi dapat menunjukkan bahwa dokumen yang secara spesifik mengarah pada web dan mobile lebih sedikit dibandingkan bidang lain. Selain itu, sebagian dokumen pengembangan aplikasi berbasis web atau mobile dapat terklasifikasi ke bidang *Software Engineering* atau Sistem Informasi apabila istilah

yang lebih dominan dalam dokumen berkaitan dengan perancangan sistem, kebutuhan sistem, database, atau pengujian perangkat lunak.

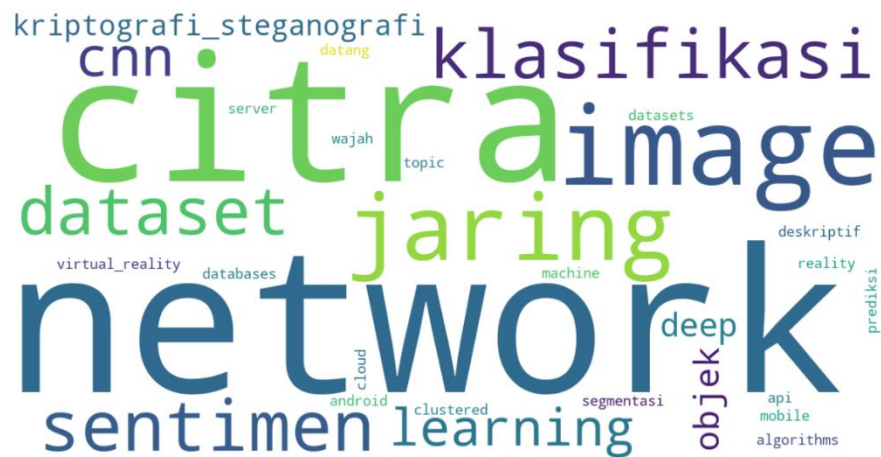
Secara keseluruhan, Gambar 4.3 menunjukkan bahwa topik skripsi mahasiswa Teknik Informatika paling banyak terkonsentrasi pada bidang *Software Engineering* dan *Intelligence System*. Kedua bidang tersebut memiliki jumlah dokumen yang hampir seimbang dan mendominasi lebih dari separuh dokumen yang berhasil dipetakan. Hal ini menunjukkan bahwa kecenderungan penelitian mahasiswa tidak hanya berfokus pada penerapan metode kecerdasan buatan, tetapi juga pada pengembangan dan rekayasa sistem perangkat lunak. Sementara itu, bidang *IoT & Robotic*, *Multimedia & Game*, *Sistem Informasi*, serta *Web & Mobile Programming* tetap memberikan kontribusi terhadap variasi tema penelitian, meskipun dengan proporsi yang lebih kecil. Dengan demikian, visualisasi ini memberikan gambaran mengenai arah dan kecenderungan penelitian mahasiswa berdasarkan bidang Teknik Informatika.

4.6.3 Visualisasi *Wordcloud* Topik

Visualisasi *Wordcloud* digunakan untuk menampilkan kata-kata dominan pada setiap topik berdasarkan frekuensi kemunculan dalam *Corpus*. Representasi ini memberikan gambaran visual mengenai karakteristik topik dengan menonjolkan kata-kata yang memiliki kontribusi paling besar. Dengan demikian, *Wordcloud* dapat membantu dalam memahami tema utama dari setiap topik secara lebih intuitif dan informatif.



Gambar 4. 4 *Wordcloud* Topik 1



Gambar 4. 5 *Wordcloud* Topik 2



bahwa model LDA mampu mengelompokkan dokumen ke dalam tema yang berbeda secara cukup jelas.

Untuk menjaga efektivitas penyajian, tidak seluruh *Wordcloud* ditampilkan dalam naskah utama, melainkan hanya beberapa topik yang representatif dan memiliki perbedaan karakteristik yang signifikan. Hal ini dilakukan untuk menghindari redundansi visual, mengingat pola distribusi kata antar topik memiliki kemiripan. Sementara itu, visualisasi *Wordcloud* untuk seluruh topik disajikan pada bagian lampiran sebagai pelengkap analisis.

4.7 Pembahasan

Tahap pembahasan dilakukan untuk menginterpretasikan hasil pemodelan topik yang telah diperoleh pada tahap sebelumnya, dengan mengacu pada hasil visualisasi, distribusi topik, serta evaluasi model. Pembahasan ini bertujuan untuk memahami makna dari pola topik yang terbentuk, kecenderungan penelitian mahasiswa, serta kualitas model yang digunakan dalam mengidentifikasi struktur topik.

4.7.1 Analisis Dominasi Topik Penelitian

Berdasarkan hasil pemetaan bidang TI, terlihat bahwa penelitian mahasiswa paling banyak mengarah pada bidang *Software Engineering* dan *Intelligence System*. *Software Engineering* memiliki jumlah 583 dokumen atau 27,89%, sedangkan *Intelligence System* memiliki 579 dokumen atau 27,70%. Selisih yang sangat kecil antara kedua bidang tersebut menunjukkan bahwa penelitian

mahasiswa tidak hanya berfokus pada penerapan kecerdasan buatan, tetapi juga pada pengembangan, perancangan, dan pengujian sistem perangkat lunak.

Dominasi *Software Engineering* menunjukkan bahwa mahasiswa banyak melakukan penelitian yang berkaitan dengan pembangunan sistem, perancangan aplikasi, analisis kebutuhan, serta evaluasi kualitas perangkat lunak. Sementara itu, dominasi *Intelligence System* menunjukkan tingginya minat terhadap metode berbasis data, seperti *machine learning*, *data mining*, klasifikasi, *clustering*, dan prediksi. Kedua bidang tersebut saling melengkapi karena pengembangan sistem modern sering kali membutuhkan integrasi antara perangkat lunak dan metode cerdas.

4.7.2 Analisis Distribusi dan Ketimpangan Topik

Distribusi bidang TI menunjukkan bahwa hasil pemetaan topik tidak tersebar secara merata. Dua bidang terbesar, yaitu *Software Engineering* dan *Intelligence System*, memiliki proporsi yang jauh lebih tinggi dibandingkan bidang lain. Sementara itu, bidang *Web & Mobile Programming* memiliki jumlah paling kecil, yaitu 105 dokumen atau 5,02%. Ketimpangan ini menunjukkan bahwa sebagian bidang lebih banyak diminati atau lebih banyak terwakili dalam dokumen skripsi dibandingkan bidang lainnya.

Ketimpangan distribusi juga dapat dipengaruhi oleh karakteristik kata dalam dokumen. Beberapa bidang seperti *Software Engineering* dan *Intelligence System* memiliki istilah yang sering muncul dalam berbagai jenis penelitian, seperti sistem, model, pengujian, klasifikasi, *data mining*, dan algoritma. Sebaliknya, bidang seperti *Web & Mobile Programming* cenderung membutuhkan istilah yang

lebih spesifik, seperti *Android*, *Laravel*, *Flutter*, *API*, atau *mobile*. Oleh karena itu, dokumen yang mengandung istilah umum pengembangan sistem dapat lebih mudah terklasifikasi ke bidang *Software Engineering* dibandingkan *Web & Mobile Programming*.

4.7.3 Analisis Hubungan Antar Topik

Hubungan antar topik dianalisis berdasarkan visualisasi pyLDAvis yang menunjukkan posisi relatif masing-masing topik dalam ruang dua dimensi. Dalam visualisasi tersebut, jarak antar topik merepresentasikan tingkat kemiripan distribusi kata, di mana topik yang berdekatan memiliki kosakata yang lebih mirip dibandingkan topik yang berjauhan.

Hasil visualisasi menunjukkan bahwa beberapa topik berada dalam posisi yang relatif berdekatan, khususnya topik yang berkaitan dengan *machine learning*, klasifikasi, dan *data mining*. Hal ini menunjukkan adanya keterkaitan semantik antar topik tersebut, yang disebabkan oleh penggunaan metode dan istilah yang serupa, seperti *learning*, klasifikasi, dan model. Kondisi ini menunjukkan bahwa meskipun topik dipisahkan, terdapat irisan konsep yang tidak dapat sepenuhnya dipisahkan.

Sebaliknya, topik yang berkaitan dengan *Internet of Things* (IoT) dan *embedded system* cenderung berada pada posisi yang lebih terpisah. Hal ini menunjukkan bahwa topik tersebut memiliki karakteristik kosakata yang berbeda, seperti *sensor*, *arduino*, dan *monitoring*. Perbedaan ini menunjukkan bahwa model LDA mampu membedakan topik berdasarkan konteks penggunaan kata, sehingga menghasilkan pemisahan topik yang cukup jelas.

Analisis ini menunjukkan bahwa model LDA tidak hanya mampu mengelompokkan dokumen, tetapi juga mampu merepresentasikan hubungan antar topik secara semantik. Hal ini menjadi salah satu indikator bahwa model yang digunakan telah bekerja dengan baik dalam menangkap struktur laten dari data teks.

4.7.4 Analisis Kualitas Model

Kualitas model dalam penelitian ini diukur menggunakan *Coherence Score*, dengan nilai terbaik sebesar 0.4439 pada jumlah topik sebanyak 9. *Coherence Score* digunakan untuk mengukur tingkat keterkaitan antar kata dalam suatu topik, sehingga dapat digunakan sebagai indikator kualitas interpretasi topik (Röder et al., 2015).

Nilai *coherence* sebesar 0.4439 menunjukkan bahwa model memiliki kualitas yang cukup baik dalam membentuk topik yang koheren dan dapat diinterpretasikan. Meskipun nilai tersebut belum mencapai kategori sangat tinggi, hasil ini sudah cukup representatif untuk data teks dengan kompleksitas tinggi seperti abstrak skripsi mahasiswa.

Namun demikian, terdapat beberapa indikasi yang menunjukkan bahwa model masih dapat ditingkatkan. Salah satunya adalah munculnya token hasil *preprocessing* yang kurang optimal, seperti pengulangan kata (*api_api*, *dijkstra_dijkstra*), yang dapat memengaruhi kualitas distribusi kata dalam topik. Selain itu, keberadaan kata umum yang masih tersisa juga dapat menyebabkan topik menjadi kurang spesifik.

Dengan demikian, kualitas model dapat dikatakan cukup baik, namun masih memiliki ruang untuk peningkatan, terutama pada tahap *preprocessing* dan pemilihan parameter model.

4.7.5 Analisis Kelebihan dan Keterbatasan Model

Model LDA yang digunakan dalam penelitian ini memiliki kelebihan dalam mengidentifikasi pola topik dari kumpulan dokumen skripsi secara otomatis. Dengan menggunakan data judul, abstrak, dan kata kunci, model mampu menghasilkan distribusi topik, kata dominan, serta visualisasi topik yang dapat digunakan untuk melihat kecenderungan penelitian mahasiswa. Selain itu, penggunaan seleksi token berbasis istilah Teknik Informatika membantu corpus menjadi lebih relevan dengan domain penelitian.

Namun, penelitian ini juga memiliki beberapa keterbatasan. Pertama, hasil model sangat bergantung pada kualitas data teks yang tersedia. Dokumen dengan abstrak yang terlalu pendek, kata kunci yang tidak lengkap, atau istilah teknis yang minim dapat gugur pada tahap seleksi data. Kedua, dari 3961 dokumen awal, hanya 3349 dokumen yang digunakan dalam corpus final LDA, sedangkan 612 dokumen gugur pada tahap filter awal dan seleksi token TI. Ketiga, dari dokumen final tersebut, sebanyak 1259 dokumen tidak terklasifikasi ke dalam enam bidang TI karena tidak memiliki kecocokan istilah yang cukup kuat terhadap kamus bidang. Hal ini menunjukkan bahwa proses pemetaan bidang masih bergantung pada kelengkapan kamus istilah dan kualitas representasi teks pada setiap dokumen. Hal ini menjadi salah satu keterbatasan umum dalam metode topic modeling berbasis probabilistik (Blei et al., 2003).

4.7.7 Implikasi Penelitian

Hasil penelitian ini memiliki implikasi yang cukup signifikan, baik dari sisi akademik maupun praktis. Dari sisi akademik, penelitian ini menunjukkan bahwa metode LDA dapat digunakan sebagai alat untuk memetakan topik penelitian secara otomatis dari kumpulan dokumen teks. Hal ini memberikan kontribusi dalam pengembangan metode analisis teks, khususnya dalam konteks pendidikan tinggi dan analisis karya ilmiah mahasiswa.

Dari sisi praktis, hasil pemetaan topik dapat digunakan sebagai referensi bagi mahasiswa dalam menentukan topik penelitian. Dengan mengetahui distribusi topik yang telah banyak diteliti, mahasiswa dapat menghindari pemilihan topik yang terlalu umum dan mencari celah penelitian yang masih jarang dikaji. Hal ini diharapkan dapat meningkatkan kualitas dan keberagaman penelitian mahasiswa.

Selain itu, hasil penelitian ini juga dapat dimanfaatkan oleh institusi pendidikan sebagai bahan evaluasi terhadap arah penelitian mahasiswa. Dengan mengetahui dominasi dan ketimpangan topik, institusi dapat merancang kebijakan atau arahan penelitian yang lebih seimbang, sehingga mendorong pengembangan bidang yang masih kurang diminati.

Lebih lanjut, temuan mengenai dominasi bidang *Software Engineering* dan *Intelligence System* menunjukkan bahwa kurikulum dan perkembangan teknologi memiliki pengaruh yang besar terhadap pilihan topik penelitian mahasiswa. Oleh karena itu, hasil ini juga dapat digunakan sebagai dasar dalam pengembangan kurikulum yang lebih adaptif terhadap perkembangan teknologi, sekaligus tetap mendorong eksplorasi pada bidang lain yang belum banyak diteliti.

4.8 Integrasi Sains dan Islam

Penelitian ini tidak hanya dipahami sebagai proses komputasi untuk mengelompokkan dokumen secara otomatis, tetapi juga sebagai bentuk ikhtiar ilmiah yang sejalan dengan nilai-nilai Islam. Dalam perspektif Islam, ilmu pengetahuan merupakan sarana untuk membaca tanda-tanda kebesaran Allah, baik yang tertulis dalam wahyu maupun yang tersaji dalam realitas kehidupan. Proses *topic modeling* terhadap kumpulan abstrak skripsi pada dasarnya merupakan usaha untuk menemukan pola, keteraturan, dan makna yang tersembunyi di dalam data teks. Kegiatan tersebut selaras dengan semangat Islam yang mendorong manusia untuk berpikir, meneliti, dan mengambil pelajaran dari setiap fenomena. Hasil penelitian ini menunjukkan bahwa metode *Latent Dirichlet Allocation* (LDA) mampu membantu peneliti mengidentifikasi kecenderungan tema penelitian mahasiswa secara sistematis. Kemampuan tersebut memperlihatkan bahwa perkembangan ilmu pengetahuan dan teknologi dapat dimanfaatkan untuk mendukung pengembangan akademik yang lebih terarah. Dengan demikian, penelitian ini dapat ditempatkan dalam kerangka integrasi sains dan Islam, yaitu memadukan metode ilmiah modern dengan pandangan hidup yang berlandaskan wahyu. Integrasi tersebut menegaskan bahwa penggunaan teknologi analisis data bukanlah sesuatu yang terpisah dari nilai keislaman, melainkan bagian dari usaha manusia dalam mengelola ilmu secara bertanggung jawab.

Landasan utama integrasi ini dapat dikaitkan dengan firman Allah Swt. dalam QS. Al-‘Alaq ayat 1–5:

أَفْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ﴿١﴾ خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ ﴿٢﴾ افْرَأْ وَرَبُّكَ الْأَكْرَمُ ﴿٣﴾ الَّذِي عَلَّمَ بِالْقَلَمِ ﴿٤﴾ عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ ﴿٥﴾

“Bacalah dengan nama Tuhanmu yang menciptakan. Dia telah menciptakan manusia dari segumpal darah. Bacalah, dan Tuhanmulah Yang Maha Mulia. Yang mengajar manusia dengan pena. Dia mengajarkan manusia apa yang tidak diketahuinya.” (Q.s Al Alaq: 1-5).

Ayat tersebut menempatkan membaca, belajar, dan menulis sebagai fondasi utama dalam tradisi keilmuan Islam. Dalam perspektif Tafsir Tarbawi, QS. Al-‘Alaq ayat 1–5 juga dipahami sebagai ayat yang menghubungkan kesadaran intelektual dengan kesadaran transendental. Wathani menjelaskan bahwa tafsir tarbawi progresif terhadap Surah Al-‘Alaq menekankan keterkaitan antara pembacaan, ketuhanan, kemanusiaan, ilmu, dan aktivitas menulis sebagai dasar pendidikan Islam (Wathani, 2025). Oleh sebab itu, penelitian ilmiah idealnya tidak hanya berorientasi pada hasil teknis, tetapi juga diarahkan untuk memberikan manfaat, memperkuat budaya literasi, dan mendukung pengembangan ilmu pengetahuan yang bertanggung jawab.

Integrasi antara sains dan Islam juga dapat dikaitkan dengan QS. Ali ‘Imran ayat 190–191:

إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَالاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ لَآيَاتٍ لِأُولِي الْأَلْبَابِ ﴿١٩٠﴾ الَّذِينَ يَذْكُرُونَ اللَّهَ قِيَامًا وَقُعُودًا وَعَلَىٰ جُنُوبِهِمْ وَيَتَفَكَّرُونَ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ رَبَّنَا مَا خَلَقْتَ هَذَا بَاطِلًا سُبْحَانَكَ فَقِنَا عَذَابَ النَّارِ ﴿١٩١﴾

“Sesungguhnya dalam penciptaan langit dan bumi, serta pergantian malam dan siang terdapat tanda-tanda bagi orang-orang yang berakal, yaitu orang-orang yang mengingat Allah sambil berdiri, duduk, dan berbaring, serta mereka memikirkan tentang penciptaan langit dan bumi seraya berkata: ‘Ya Tuhan kami, tidaklah Engkau menciptakan semua ini sia-sia. Mahasuci Engkau, maka lindungilah kami dari azab neraka.’” (QS. Ali ‘Imran ayat 190–191).

Menurut Tafsir Al-Misbah, ayat ini menunjukkan pentingnya keseimbangan antara dzikir dan pikir. Dzikir menggambarkan hubungan spiritual manusia dengan Allah, sedangkan pikir menunjukkan penggunaan akal dalam memahami fenomena kehidupan dan alam semesta. Quraish Shihab menjelaskan bahwa Islam mendorong umatnya untuk melakukan pengamatan, penelitian, dan perenungan terhadap berbagai fenomena sebagai jalan untuk memperkuat keimanan dan memperluas ilmu pengetahuan (Shihab, 2002).

Sementara itu, dalam Tafsir Al-Azhar dijelaskan bahwa ayat ini merupakan dorongan bagi manusia untuk mengembangkan ilmu pengetahuan melalui pengamatan terhadap alam dan kehidupan. Hamka menegaskan bahwa orang beriman tidak hanya cukup beribadah secara ritual, tetapi juga harus menggunakan akalnya untuk menelaah berbagai tanda kebesaran Allah. Dengan berpikir dan meneliti, manusia akan memahami bahwa ilmu pengetahuan merupakan sarana untuk mengenal kekuasaan Allah serta membangun kemaslahatan bagi kehidupan manusia (Hamka, 1982).

Pemanfaatan metode LDA dalam penelitian ini juga mencerminkan pentingnya akal sebagai anugerah Allah yang harus digunakan secara optimal. Islam tidak memandang akal sebagai lawan wahyu, melainkan sebagai instrumen untuk memahami ayat-ayat Allah, baik yang bersifat qauliyah maupun kauniyah. Ketika peneliti melakukan *preprocessing*, membangun *Corpus*, melatih model, dan mengevaluasi *Coherence Score*, seluruh tahapan tersebut merupakan bentuk kerja akal yang sistematis. Proses itu menuntut ketelitian, logika, dan kemampuan analisis agar hasil yang diperoleh tidak keliru. Dalam Islam, penggunaan akal

secara benar merupakan bagian dari tanggung jawab manusia sebagai khalifah di muka bumi. Melalui penelitian ini, terlihat bahwa akal manusia dapat dipadukan dengan teknologi komputasi untuk membantu memetakan tren topik skripsi secara lebih objektif. Oleh sebab itu, keberhasilan model dalam menemukan enam topik utama dapat dipahami sebagai hasil dari penggunaan akal yang diarahkan pada pengembangan ilmu yang bermanfaat.

Ayat lain yang sangat relevan adalah QS. Az-Zumar ayat 9:

قُلْ هَلْ يَسْتَوِي الَّذِينَ يَعْلَمُونَ وَالَّذِينَ لَا يَعْلَمُونَ ۚ إِنَّمَا يَتَذَكَّرُ أُولُو الْأَلْبَابِ

“Katakanlah, apakah sama orang-orang yang mengetahui dengan orang-orang yang tidak mengetahui? Sesungguhnya hanya orang yang berakal sehat yang dapat menerima pelajaran.” (QS. Az-Zumar ayat 9).

Secara tafsir, ayat ini menegaskan kemuliaan orang berilmu, karena ilmu menjadikan manusia mampu membedakan yang benar dan yang tidak benar, yang bermanfaat dan yang tidak bermanfaat. Dalam konteks penelitian ini, analisis *topic mining* membantu menghadirkan pengetahuan baru dari kumpulan data skripsi yang sangat banyak dan sulit dibaca secara manual.

Dalam Tafsir Al-Misbah dijelaskan bahwa ayat ini menunjukkan kemuliaan ilmu pengetahuan dalam Islam. Quraish Shihab menerangkan bahwa pertanyaan “apakah sama orang-orang yang mengetahui dengan orang-orang yang tidak mengetahui?” merupakan bentuk penegasan bahwa ilmu memiliki kedudukan yang sangat tinggi dalam kehidupan manusia. Orang yang memiliki ilmu akan lebih mudah memahami hakikat kehidupan, mengambil keputusan, dan membedakan antara yang benar dan yang salah (Shihab, 2002).

Dari hasil penelitian juga ditemukan bahwa salah satu topik masih didominasi oleh kata-kata umum berbahasa Inggris, sehingga menunjukkan adanya *noise* linguistik yang belum sepenuhnya tersaring pada tahap *preprocessing*. Temuan ini dapat dikaitkan dengan prinsip *tabayyun* dalam Islam, yaitu sikap memeriksa, memverifikasi, dan meneliti informasi sebelum menarik kesimpulan. Hal tersebut sejalan dengan QS. Al-Hujurat ayat 6:

يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ

“Wahai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka telitilah kebenarannya, agar kamu tidak mencelakakan suatu kaum karena kebodohan, yang akhirnya membuat kamu menyesali perbuatanmu itu.” (QS. Al-Hujurat ayat 6).

Menurut Tafsir Ibnu Katsir, QS. Al-Hujurat ayat 6 turun sebagai peringatan agar umat Islam tidak langsung mempercayai suatu informasi tanpa melakukan pemeriksaan terlebih dahulu. Ibnu Katsir menjelaskan bahwa apabila berita diterima dari orang yang tidak jelas kebenarannya, maka informasi tersebut harus diverifikasi agar tidak menimbulkan kesalahan, fitnah, maupun kerugian bagi pihak lain. Ayat ini menjadi dasar penting dalam Islam mengenai prinsip kehati-hatian, *tabayyun*, dan validasi informasi sebelum mengambil keputusan (Ibnu Katsir, 2004).

Dalam penelitian ini, *preprocessing* dapat dipahami sebagai bentuk *tabayyun* terhadap data, karena peneliti harus menyaring kata-kata yang tidak relevan agar model menghasilkan topik yang lebih akurat. Ketika masih muncul *noise* bahasa Inggris, hal itu menunjukkan bahwa proses verifikasi data masih dapat

ditingkatkan. Dengan demikian, aspek teknis dalam pengolahan teks ternyata juga memiliki nilai etis yang sejalan dengan ajaran Islam tentang kehati-hatian dalam mengelola informasi. Integrasi ini menunjukkan bahwa kualitas hasil ilmiah sangat bergantung pada kejujuran, ketelitian, dan tanggung jawab peneliti dalam memproses data.

Hasil optimasi model yang menunjukkan bahwa jumlah topik terbaik adalah enam berdasarkan nilai *Coherence Score* juga dapat dipahami dalam perspektif *itqan*, yaitu bekerja secara sungguh-sungguh, teliti, dan menghasilkan kualitas terbaik. Peneliti tidak langsung menetapkan jumlah topik secara sembarangan, tetapi melakukan pengujian dari berbagai kemungkinan hingga diperoleh konfigurasi terbaik. Tindakan tersebut mencerminkan semangat profesionalisme ilmiah yang sangat dekat dengan nilai *itqan* dalam Islam. Secara substantif, pemilihan model terbaik menunjukkan bahwa penelitian dilakukan dengan standar evaluasi yang terukur, bukan berdasarkan dugaan subjektif. Islam mendorong setiap pekerjaan dilakukan secara optimal agar hasilnya benar-benar memberi manfaat. Ketika model dengan enam topik menghasilkan *Coherence Score* tertinggi, keputusan tersebut menunjukkan usaha memilih hasil yang paling tepat dari beberapa alternatif yang tersedia. Oleh karena itu, proses optimasi model dalam penelitian ini tidak hanya merupakan prosedur teknis, tetapi juga mencerminkan etos kerja ilmiah yang sejalan dengan ajaran Islam.

Integrasi sains dan Islam dalam penelitian ini juga tampak pada tujuan praktisnya, yaitu membantu mahasiswa menemukan arah topik penelitian yang relevan dan bermanfaat. Dalam Islam, ilmu yang terbaik adalah ilmu yang memberi

manfaat bagi orang lain. Hasil penelitian ini berupa peta topik yang dapat digunakan sebagai referensi awal dalam memahami kecenderungan penelitian skripsi. Dengan adanya peta tersebut, mahasiswa dapat lebih mudah melihat bidang yang sudah banyak diteliti maupun bidang yang masih memiliki peluang pengembangan. Hal ini membuka ruang bagi lahirnya penelitian yang lebih inovatif, terarah, dan tidak hanya mengulang tema-tema sebelumnya. Dari sudut pandang Islam, upaya mempermudah akses terhadap ilmu dan memperbaiki kualitas pendidikan merupakan bentuk amal yang bernilai kebaikan. Dengan demikian, teknologi analisis teks dalam penelitian ini tidak berhenti sebagai alat hitung semata, tetapi menjadi sarana pelayanan akademik yang lebih luas.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode Latent Dirichlet Allocation (LDA) dapat digunakan untuk mengidentifikasi dan memetakan tren topik skripsi mahasiswa berdasarkan data judul, abstrak, dan kata kunci.

Berdasarkan hasil pemetaan ke enam bidang Teknik Informatika, sebanyak 2090 dokumen berhasil dipetakan ke dalam bidang TI. Bidang dengan jumlah tertinggi adalah *Software Engineering* sebanyak 583 dokumen, diikuti oleh *Intelligence System* sebanyak 579 dokumen. Selanjutnya, bidang *IoT & Robotic* memiliki 359 dokumen, *Multimedia & Game* memiliki 281 dokumen, Sistem Informasi memiliki 183 dokumen, dan *Web & Mobile Programming* memiliki 105 dokumen. Hasil ini menunjukkan bahwa tren penelitian mahasiswa paling banyak mengarah pada pengembangan perangkat lunak dan sistem cerdas.

Dengan demikian, penelitian ini berhasil menghasilkan pemetaan tren topik skripsi mahasiswa menggunakan pendekatan *topic mining* berbasis LDA. Hasil penelitian dapat digunakan sebagai gambaran mengenai arah penelitian mahasiswa Teknik Informatika serta menjadi referensi awal bagi mahasiswa dalam mengeksplorasi topik skripsi yang relevan dengan bidang keilmuan.

5.2 Saran

- a) Pada penelitian ini terdapat sejumlah data yang gugur pada tahap seleksi, yaitu 406 dokumen gugur pada filter awal domain TI dan 206 dokumen gugur pada seleksi token TI. Oleh karena itu, penelitian selanjutnya disarankan melakukan perbaikan kualitas dataset sebelum pemodelan, misalnya dengan melengkapi abstrak yang kosong, memperbaiki kata kunci yang terlalu umum, serta memastikan bahwa setiap dokumen memiliki istilah teknis yang cukup sesuai dengan bidang Teknik Informatika.
- b) Penelitian selanjutnya dapat mengembangkan kamus istilah bidang Teknik Informatika yang lebih lengkap dan tervalidasi oleh ahli atau dosen bidang terkait. Hal ini diperlukan karena pada penelitian ini masih terdapat 1259 dokumen yang tidak terklasifikasi ke dalam enam bidang TI. Dengan kamus istilah yang lebih kaya, proses pemetaan bidang dapat menjadi lebih akurat dan mengurangi jumlah dokumen tidak terklasifikasi.
- c) Penelitian selanjutnya dapat menggunakan isi lengkap dokumen skripsi, tidak hanya judul, abstrak, dan kata kunci.
- d) Hasil pemetaan topik pada penelitian ini dapat dikembangkan menjadi sistem rekomendasi topik skripsi. Sistem tersebut dapat membantu mahasiswa dalam mengeksplorasi topik penelitian yang relevan, melihat tren bidang yang sedang berkembang, serta menemukan peluang penelitian yang masih jarang dibahas.

DAFTAR PUSTAKA

- Asmussen, C. B., & Møller, C. (2019). *Smart literature review: a practical topic modelling approach to exploratory literature review*. *Journal of Big Data*, 6(1). <https://doi.org/10.1186/s40537-019-0255-7>
- Atmaja, D. M. U., & Mandala, R. (2020). Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan *Vector Space Model* dan Metode *K-Nearest Neighbor*. *IT for Society*, 4(2), 1–6. <https://doi.org/10.33021/itfs.v4i2.1182>
- Cai, Y., Wu, X., Xie, X., & Xu, J. (2019). *A topic mining method for multi-source network public opinion based on improved hierarchical clustering*. *Proceedings - 2019 IEEE 4th International Conference on Data Science in Cyberspace, DSC 2019*, 439–444. <https://doi.org/10.1109/DSC.2019.00073>
- Nugraha, E., Elektro, T., & Elektro, T. (2024). *Topic Modeling Dokumen Skripsi Prodi Pendidikan Universitas Negeri Jakarta Menggunakan Available At*: 8(2), 21–30.
- Rashid, J., Adnan Shah, S. M., Irtaza, A., Mahmood, T., Nisar, M. W., Shafiq, M., & Gardezi, A. (2019). *Topic Modeling Technique for Text Mining over Biomedical Text Corpora through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering*. *IEEE Access*, 7, 146070–146080. <https://doi.org/10.1109/ACCESS.2019.2944973>
- Shah, A. M., Yan, X., Qayyum, A., Naqvi, R. A., & Shah, S. J. (2021). *Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach*. *International Journal of Medical Informatics*, 149(July 2020), 104434. <https://doi.org/10.1016/j.ijmedinf.2021.104434>
- Sun, J., Li, L., Li, W., Zhang, J., & Yan, C. (2020). *Enabling 5G: sentimental image dominant graph topic model for cross-modality topic detection*. *Wireless Networks*, 26(3), 1549–1561. <https://doi.org/10.1007/s11276-019-02009-3>
- Zhang, T., Lee, B., Zhu, Q., Han, X., & Ye, E. M. (2020). *Multi-Dimension Topic mining Based on Hierarchical Semantic Graph Model*. *IEEE Access*, 8, 64820–64835. <https://doi.org/10.1109/ACCESS.2020.2984352>
- Sabirin, F., Sulistiyarini, D., & Zulkarnain, Z. (2020). Pengembangan Sistem Informasi Seminar dan Skripsi Mahasiswa. *Edumatic : Jurnal Pendidikan Informatika*, 4(1), 73–82. <https://doi.org/10.29408/edumatic.v4i1.2048>
- Baliyan, A., Kaswan, K. S., & Dhatteerwal, J. S. (2022). *An Empirical Analysis of Python Programming for Advance Computing*. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1482–1486.

<https://doi.org/10.1109/ICACITE53722.2022.9823643>

- Chai, C. P. (2023). Comparison of text *preprocessing* methods. *Natural Language Engineering*, 29(3), 509–553. <https://doi.org/10.1017/S1351324922000213>
- Hadi, N. F., & Afandi, N. K. (2021). Literature Review is A Part of Research. *Sultra Educational Journal*, 1(3), 64–71. <https://doi.org/10.54297/seduj.v1i3.203>
- Ismanto, H. S., Wakhjudin, H., & Baedowi, S. (2023). Analisis Kualitatif Tingkat Stres Mahasiswa Dalam Menyelesaikan Skripsi Di Fakultas Ilmu Pendidikan Universitas PGRI Semarang. *Wawasan Pendidikan*, 3(2), 510–516. <https://doi.org/10.26877/wp.v3i2.15355>
- Meng, F., Qi, M., & Luo, C. (2024). Research Progress of Review *Topic mining* Methods: From Word Frequency Statistics to Deep Learning. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 737–741. <https://doi.org/10.1109/CSNT60213.2024.10545711>
- Nava, J., & Lee, H. (2023). mpiPython: Prospects for Node Performance. *2023 6th International Conference on Information and Computer Technologies (ICICT)*, 182–187. <https://doi.org/10.1109/ICICT58900.2023.00038>
- Nurcaya. (2021). Berpikir Kritis dalam Skripsi Mahasiswa. *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, 7(1), 106–111. <https://doi.org/10.30605/onoma.v7i1.572>
- Rajagopalan, G. (2021). Getting Familiar with Python. In *A Python Data Analyst's Toolkit* (pp. 1–43). Apress. https://doi.org/10.1007/978-1-4842-6399-0_1
- Septiono, T. (2021). Alumni Reflection To The Experience Of Writing Thesis: Case Study Of Alumni Program In Library And Information Studies Brawijaya University. *JPUA: Jurnal Perpustakaan Universitas Airlangga: Media Informasi Dan Komunikasi Kepustakawanan*, 11(2), 101. <https://doi.org/10.20473/jpua.v11i2.2021.101-112>
- Sharma, D., Yadav, H., Pandey, N., & Pandey, V. (2025). A Review Paper on Python. *International Journal of Innovative Research in Engineering*, 119–121. <https://doi.org/10.59256/ijire.20250602014>
- Sidiq, A. (2024). *Identifikasi Isu Strategis Tangerang Selatan Melalui Analisis Perbincangan di Media Sosial dengan Latent Dirichlet Allocation*. <http://repository.ipb.ac.id/handle/123456789/147242>
- Subeno, B., Kusumaningrum, R., & Fakhirin. (n.d.). *OPTIMASI JUMLAH TOPIK KORPUS MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA)*.
- Alattar, F., & Shaalan, K. (2021). *Emerging research topic detection using*

- Filtered-LDA*. *AI*, 2(4), 578–599. <https://doi.org/10.3390/ai2040035>.
- 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>.
- Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787–2805. <https://doi.org/10.1016/j.comnet.2010.05.010>.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Booth, W. C., Colomb, G. G., Williams, J. M., Bizup, J., & FitzGerald, W. T. (2016). *The craft of research* (4th ed.). The University of Chicago Press.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211–244.
- Hart, C. (2018). *Doing a literature review: Releasing the research imagination*. SAGE Publications.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>.
- Mukmin, T. (2016). Urgensi belajar dalam perspektif Al-Qur'an surat Al-'Alaq ayat 1–5 menurut Tafsir Ibnu Katsir. *El-Ghiroh*, 11(2), 1–21.
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. <https://doi.org/10.1145/2684822.2685324>.
- Setijohatmo, U. T., Rachmat, S., Susilawati, T., & Rahman, Y. (2020). Analisis metoda Latent Dirichlet Allocation untuk klasifikasi dokumen laporan tugas akhir berdasarkan pemodelan topik. *Prosiding Industrial Research Workshop and National Seminar*, 11(1). <https://doi.org/10.35313/irwns.v11i1.2040>.
- Steyvers, M., & Griffiths, T. (2007). *Probabilistic topic models*. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.
- Syed, S., & Spruit, M. (2017). *Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation*. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. <https://doi.org/10.1109/DSAA.2017.61>.
- Zimmermann, J., Champagne, L. E., Dickens, J. M., & Hazen, B. T. (2024).

Approaches to improve *preprocessing* for Latent Dirichlet Allocation topic modeling. *Decision Support Systems*.

Syaifullah Wathani. 2025. “Mengembangkan Tafsir Maqāṣidiy dari Maqāṣid Al-Qur’ān: Tafsir Tarbawi Progresif atas Surah Al-‘Alaq (96): 1–5.” *Ṣuḥuf*, Vol. 18, No. 1, hlm. 27–50. DOI: 10.22548/shf.v18i1.1286.

M. Quraish Shihab. 2002. *Tafsir Al-Misbah: Pesan, Kesan, dan Keserasian Al-Qur’an*. Jakarta: Lentera Hati.

Hamka. 1982. *Tafsir Al-Azhar*. Jakarta: Pustaka Panjimas.

Ibnu Katsir. 2004. *Tafsir Al-Qur’an Al-‘Azhim*. Beirut: Dar Al-Kutub Al-‘Ilmiyyah.