

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan ilmu pengetahuan dan teknologi di bidang Teknik Informatika berlangsung sangat pesat dan dinamis. Setiap tahunnya, mahasiswa program sarjana menghasilkan karya ilmiah berupa skripsi yang mencerminkan topik dan arah penelitian terkini. Topik-topik seperti kecerdasan buatan, *data mining*, keamanan informasi, pengolahan citra, serta pengembangan perangkat lunak terus mengalami perkembangan seiring dengan kemajuan teknologi dan kebutuhan industri (Bahrani & Fathurrahmad, 2020; Gupta et al., 2022; Hancar et al., 2025). Hal ini menunjukkan bahwa skripsi tidak hanya berfungsi sebagai syarat akademik, tetapi juga merepresentasikan tren penelitian aktual dalam suatu bidang keilmuan. Dengan demikian, kumpulan abstrak skripsi dapat menjadi sumber data yang berharga untuk memahami pola dan arah perkembangan penelitian dari waktu ke waktu.

Seiring dengan meningkatnya jumlah publikasi skripsi, volume data abstrak yang tersedia juga semakin besar dan beragam. Di Indonesia, puluhan ribu abstrak skripsi Teknik Informatika dihasilkan setiap tahun dari berbagai perguruan tinggi. Kondisi ini menimbulkan tantangan dalam proses pengelompokan dokumen penelitian karena setiap abstrak dapat menggunakan istilah, metode, dan konteks yang berbeda meskipun membahas tema yang serupa. Pengelompokan secara manual tidak hanya membutuhkan waktu yang besar, tetapi juga rentan terhadap subjektivitas dalam menentukan kesamaan antar dokumen. Selain itu, pendekatan berbasis pencocokan kata (*keyword matching*) sering kali kurang mampu menangkap hubungan makna yang terkandung dalam teks, sehingga diperlukan metode klusterisasi yang mampu merepresentasikan kemiripan dokumen berdasarkan konteks semantik, bukan hanya berdasarkan kemunculan kata yang sama.

Abstrak skripsi merupakan sumber data tekstual yang kaya informasi karena memuat ringkasan tujuan, metode, serta hasil penelitian. Dalam konteks ini, pendekatan *text mining* dan *Natural Language Processing* (NLP) dapat dimanfaatkan untuk mengekstraksi pola dan informasi tersembunyi dari data teks tidak terstruktur (Abdelrazek et al., 2023; Khurana et al., 2023). Salah satu teknik yang banyak digunakan adalah *topic modeling*, yaitu metode *unsupervised learning* yang bertujuan untuk mengidentifikasi struktur tema laten dalam kumpulan dokumen (Jelodar et al., 2019; Zhao et al., 2021). Metode klasik seperti *Latent Dirichlet Allocation* (LDA) telah banyak digunakan dalam berbagai penelitian untuk analisis tren topik (Blei et al., 2003; Gupta et al., 2022; Ozyurt et al., 2024). Namun, pendekatan berbasis *bag-of-words* pada LDA memiliki keterbatasan dalam menangkap konteks semantik, terutama pada teks pendek dan variatif seperti abstrak skripsi (Muthusami et al., 2024; Zhao et al., 2021).

Klasterisasi dokumen teks merupakan salah satu pendekatan *unsupervised learning* yang bertujuan mengelompokkan dokumen berdasarkan tingkat kemiripan karakteristik yang dimiliki. Dalam konteks abstrak skripsi, proses klasterisasi menjadi lebih menantang karena dokumen tidak hanya mengandung kata-kata yang berbeda, tetapi juga makna yang saling berkaitan secara kontekstual. Oleh karena itu, diperlukan pendekatan yang mampu merepresentasikan informasi semantik dokumen secara lebih baik agar abstrak dengan tema serupa dapat ditempatkan pada klaster yang sama meskipun menggunakan kosakata yang berbeda. Hasil klaster yang terbentuk selanjutnya dapat diinterpretasikan melalui kata kunci dominan untuk memperoleh gambaran topik penelitian yang muncul dalam kumpulan dokumen.

Perkembangan model bahasa berbasis *transformer*, seperti BERT (Devlin et al., 2019) dan Sentence-BERT (Reimers & Gurevych, 2019), memberikan pendekatan baru dalam representasi teks yang mampu memahami konteks semantik secara lebih mendalam. Pendekatan ini memungkinkan representasi dokumen dalam bentuk *embedding* yang mempertimbangkan hubungan makna antar kata dalam suatu kalimat. Salah satu metode *topic modeling* yang memanfaatkan keunggulan tersebut adalah BERTopic yang dikembangkan oleh Grootendorst (2022).

BERTopic mengombinasikan *transformer-based embedding*, reduksi dimensi menggunakan UMAP (McInnes et al., 2018), klasterisasi dengan HDBSCAN (McInnes et al., 2017), serta ekstraksi kata kunci berbasis *class-based TF-IDF* (c-TF-IDF). Dibandingkan dengan metode tradisional seperti LDA, pendekatan BERTopic memiliki keunggulan dalam menangkap konteks semantik, terutama pada data teks pendek yang memiliki keterbatasan informasi kontekstual (Muthusami et al., 2024; Zhao et al., 2021). Penelitian oleh Egger dan Yu (2022) menunjukkan bahwa BERTopic menghasilkan koherensi topik yang lebih tinggi dibandingkan LDA, NMF, dan Top2Vec pada data teks pendek. Karakteristik ini sejalan dengan data dalam penelitian ini, yaitu abstrak skripsi yang relatif singkat, padat, dan memiliki variasi topik yang tinggi, sehingga memerlukan metode yang mampu menangkap hubungan semantik secara kontekstual.

Selain itu, BERTopic tidak memerlukan penentuan jumlah topik secara apriori karena proses klasterisasi dilakukan secara otomatis menggunakan HDBSCAN (Grootendorst, 2022), sehingga lebih fleksibel dalam menangani data dengan distribusi topik yang tidak diketahui sebelumnya. Metode ini juga mendukung analisis temporal (*topic evolution*), yang memungkinkan identifikasi perubahan dan perkembangan topik dari waktu ke waktu (Grootendorst, 2022; Raman et al., 2024). Fitur ini menjadi relevan dengan tujuan penelitian yang tidak hanya mengidentifikasi topik, tetapi juga menganalisis evolusi tren penelitian skripsi berdasarkan dimensi waktu.

Tantangan utama yang dihadapi dalam penelitian ini tidak hanya terletak pada kemampuan membentuk klaster secara otomatis, tetapi juga pada kemampuan menghasilkan klaster yang memiliki kualitas baik dan dapat dipertanggungjawabkan secara ilmiah. Klaster yang terbentuk harus memiliki koherensi tematik yang tinggi sehingga kata-kata dalam satu topik saling berkaitan secara semantik, serta memiliki keberagaman yang memadai agar setiap topik merepresentasikan tema yang berbeda. Tanpa ukuran keberhasilan yang jelas, hasil klasterisasi tidak dapat dipertanggungjawabkan secara ilmiah. Oleh karena itu, diperlukan metrik evaluasi yang mampu mengukur kualitas topik secara objektif. Dalam penelitian ini, dua metrik internal digunakan: *Topic Coherence* (Cv) untuk

mengukur koherensi semantik antar kata kunci dalam satu klaster, dan *Topic Diversity* (TD) untuk mengukur keberagaman topik antar klaster. Dengan metrik tersebut, konfigurasi parameter BERTopic yang menghasilkan nilai Cv dan TD terbaik dapat dipilih sebagai model yang paling valid untuk analisis selanjutnya.

Berdasarkan kajian terhadap penelitian terdahulu, diketahui bahwa metode BERTopic telah terbukti mampu menghasilkan topik yang lebih koheren dibandingkan metode tradisional seperti LDA, khususnya pada data teks pendek. Namun, sebagian besar penelitian masih terbatas pada data non-ilmiah seperti media sosial atau artikel jurnal berbahasa Inggris, sehingga belum merepresentasikan karakteristik teks ilmiah pendek dalam bahasa Indonesia, khususnya abstrak skripsi. Penelitian yang secara khusus mengintegrasikan eksplorasi parameter BERTopic, evaluasi kualitas topik menggunakan Cv dan TD secara simultan, serta analisis evolusi topik temporal pada korpus abstrak skripsi Indonesia masih belum banyak dilakukan. Kesenjangan inilah yang menjadi motivasi utama penelitian ini.

Dengan mempertimbangkan permasalahan, karakteristik data, serta kesenjangan pada penelitian sebelumnya, penelitian ini bertujuan untuk melakukan pemodelan topik secara otomatis terhadap abstrak skripsi Teknik Informatika menggunakan metode BERTopic, dengan menentukan konfigurasi parameter terbaik berdasarkan evaluasi Cv dan TD. Selanjutnya, topik-topik yang dihasilkan diinterpretasikan secara deskriptif melalui kata kunci dominan dan dokumen representatif, kemudian dianalisis distribusi temporalnya untuk mengidentifikasi tren penelitian yang meningkat (*emerging*) maupun menurun (*declining*). Hasil penelitian disajikan dalam bentuk visualisasi interaktif untuk memudahkan eksplorasi. Penelitian ini diharapkan dapat memberikan gambaran yang lebih objektif mengenai pola dan kecenderungan penelitian skripsi Teknik Informatika, serta menjadi referensi bagi mahasiswa dalam melihat variasi topik yang pernah diteliti.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Perkembangan penelitian di bidang Teknik Informatika ditandai dengan meningkatnya jumlah skripsi yang dihasilkan mahasiswa setiap tahunnya. Kumpulan abstrak skripsi dari berbagai perguruan tinggi di Indonesia telah mencapai puluhan ribu dokumen dan terus bertambah seiring waktu. Data ini menyimpan informasi berharga mengenai topik-topik penelitian yang pernah dilakukan, metode yang digunakan, serta kecenderungan arah riset dari tahun ke tahun. Namun, volume data yang besar dan terus bertumbuh menyebabkan proses pemetaan topik penelitian secara manual menjadi tidak efektif dan tidak efisien. Pengguna—baik mahasiswa, dosen, maupun pengelola program studi—menghadapi kesulitan dalam mengidentifikasi pola dan tren penelitian yang tersembunyi di balik tumpukan abstrak tersebut.

Pendekatan berbasis pencocokan kata kunci (*keyword matching*) yang umum digunakan juga memiliki keterbatasan karena tidak mampu menangkap hubungan semantik antar dokumen. Dua abstrak yang membahas tema serupa tetapi menggunakan istilah yang berbeda dapat dianggap tidak memiliki keterkaitan. Oleh karena itu, diperlukan pendekatan otomatis berbasis *topic modeling* yang mampu mengekstraksi topik-topik penelitian berdasarkan kemiripan makna, bukan sekadar kemiripan kata.

Salah satu metode *topic modeling* modern yang menjanjikan adalah BERTopic, yang menggabungkan representasi semantik berbasis *transformer* dengan klasterisasi berbasis kepadatan. Namun demikian, penerapan BERTopic pada korpus abstrak skripsi berbahasa Indonesia menghadapi tantangan tersendiri. Jumlah dan kualitas klaster yang dihasilkan sangat bergantung pada konfigurasi parameter pada komponen UMAP dan HDBSCAN. Tanpa eksplorasi parameter yang sistematis, hasil klasterisasi dapat menjadi tidak optimal—baik karena terlalu banyak klaster sehingga topik menjadi terfragmentasi, maupun terlalu sedikit sehingga topik yang berbeda tercampur menjadi satu. Selain itu, kualitas topik yang

dihasilkan belum tentu koheren dan beragam secara semantik, sehingga diperlukan metrik evaluasi yang objektif untuk menilai dan memilih konfigurasi terbaik.

Selain aspek teknis pemodelan, urgensi lain dari penelitian ini adalah kebutuhan untuk melihat dinamika perubahan topik dari waktu ke waktu. Dengan ribuan abstrak yang tersebar dalam rentang waktu 12 tahun, belum diketahui secara jelas topik-topik mana yang mengalami peningkatan proporsi (*emerging*) maupun penurunan (*declining*). Informasi tren ini sangat dibutuhkan untuk memberikan gambaran arah perkembangan penelitian Teknik Informatika di Indonesia, namun belum tersedia secara terstruktur

Berdasarkan uraian tersebut, permasalahan utama yang dihadapi dalam penelitian ini adalah:

1. Belum diketahui konfigurasi parameter *fine-tuning* BERTopic yang optimal untuk menghasilkan kualitas topik terbaik pada korpus abstrak skripsi Teknik Informatika, karena parameter UMAP dan HDBSCAN sangat mempengaruhi jumlah dan koherensi klaster yang terbentuk.
2. Belum diketahui jumlah klaster optimal yang dihasilkan dari konfigurasi parameter terbaik, sehingga belum ada kepastian mengenai berapa banyak topik penelitian yang secara alami terbentuk dari data abstrak.
3. Belum diketahui pola tren temporal topik-topik penelitian Teknik Informatika selama periode 2013–2025, termasuk identifikasi topik mana yang menunjukkan kecenderungan meningkat (*emerging*) dan menurun (*declining*).

1.2.2 Metode Usulan

Penelitian ini mengusulkan penerapan BERTopic untuk melakukan klasterisasi semantik terhadap abstrak skripsi Teknik Informatika, dengan pendekatan *fine-tuning* parameter sistematis dua tahap (*Coarse Search* dan *OFAT*). Secara garis besar, metode usulan terdiri atas tahapan berikut:

1. Preprocessing teks mencakup *case folding*, pembersihan teks, tokenisasi, *stopword removal*, *stemming*, serta normalisasi istilah teknis.
2. Pembentukan *sentence embedding* menggunakan model *paraphrase-multilingual-MiniLM*.
3. Reduksi dimensi dengan UMAP.
4. Klasterisasi dengan HDBSCAN.
5. Ekstraksi kata kunci dengan c-TF-IDF.
6. Interpretasi dan pelabelan topik dilakukan oleh peneliti berdasarkan kata kunci dominan dan dokumen representatif dari setiap klaster. Label yang diberikan didiskusikan dengan dosen pembimbing untuk memastikan kesesuaian makna. Proses ini bersifat deskriptif dan bukan merupakan klasifikasi formal. Tidak digunakan skor numerik untuk menilai validitas label; konfirmasi dilakukan secara kualitatif.
7. Eksplorasi parameter pada UMAP (*n_neighbors*, *min_dist*) dan HDBSCAN (*min_cluster_size*) untuk menemukan konfigurasi terbaik.
8. Evaluasi kualitas topik menggunakan Topic Coherence (Cv) dan Topic Diversity (TD). Model terbaik dipilih berdasarkan kombinasi nilai Cv dan TD tertinggi dibandingkan skenario lain, bukan berdasarkan ambang batas absolut.
9. Analisis tren temporal dengan menghitung distribusi topik per tahun publikasi, kemudian mengklasifikasikan topik ke dalam kategori *emerging* (meningkat), *declining* (menurun), atau stabil.
10. Visualisasi interaktif untuk eksplorasi hasil.

Dalam proses interpretasi, peneliti dapat merujuk pada bidang kajian umum Teknik Informatika (termasuk referensi seperti APTIKOM) sebagai panduan

deskriptif untuk meningkatkan konsistensi pelabelan topik. Referensi tersebut tidak digunakan sebagai *ground truth* maupun klasifikasi formal.

1.2.3 Pertanyaan Penelitian

Berdasarkan permasalahan yang telah diuraikan, maka pertanyaan penelitian dalam penelitian ini adalah:

1. Bagaimana menentukan konfigurasi parameter *fine-tuning* BERTopic yang optimal—terutama pada komponen UMAP ($n_neighbors$, min_dist) dan HDBSCAN ($min_cluster_size$, $min_samples$) untuk menghasilkan pemodelan topik terbaik pada korpus abstrak skripsi Teknik Informatika, berdasarkan evaluasi kualitas topik menggunakan *Topic Coherence* (Cv) dan *Topic Diversity* (TD)?
2. Berapa jumlah klaster optimal yang dihasilkan dari konfigurasi parameter BERTopic terbaik, dan bagaimana karakteristik topik-topik penelitian yang teridentifikasi dari klaster tersebut pada korpus abstrak skripsi Teknik Informatika?
3. Bagaimana pola tren temporal topik-topik penelitian Teknik Informatika selama periode 2013–2025 berdasarkan distribusi proporsi topik per tahun, khususnya dalam mengidentifikasi topik-topik yang menunjukkan kecenderungan meningkat (*emerging*) dan menurun (*declining*)?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

1. Menentukan konfigurasi parameter BERTopic yang menghasilkan kualitas topik terbaik berdasarkan nilai *Topic Coherence* (Cv) dan *Topic Diversity* (TD) pada korpus abstrak skripsi Teknik Informatika.
2. Mengidentifikasi dan menganalisis karakteristik topik penelitian dominan yang terbentuk dari hasil klasterisasi semantik abstrak skripsi Teknik Informatika menggunakan BERTopic, melalui interpretasi kata

kunci dan dokumen representatif yang dikonfirmasi kepada dosen pembimbing.

3. Menganalisis pola perkembangan topik penelitian Teknik Informatika selama periode 2013–2025, termasuk mengidentifikasi topik yang menunjukkan tren meningkat (*emerging*) maupun menurun (*declining*) berdasarkan distribusi temporal.

1.3.2 Manfaat

- Manfaat Teoritis
 1. Memberikan kontribusi dalam penerapan BERTopic untuk analisis topik pada korpus abstrak skripsi berbahasa Indonesia, yang masih jarang dilakukan.
 2. Menjadi referensi bagi penelitian selanjutnya yang memanfaatkan *topic modeling* berbasis *transformer* untuk analisis tren penelitian.
 3. Menambah kajian mengenai evaluasi kualitas topik menggunakan *Topic Coherence* (Cv) dan *Topic Diversity* (TD) pada teks pendek ilmiah.
- Manfaat Praktis
 1. Bagi mahasiswa: Memberikan gambaran mengenai variasi topik dan kecenderungan penelitian yang pernah dilakukan, sehingga dapat membantu dalam mengeksplorasi topik skripsi yang relevan.
 2. Bagi program studi / institusi: Menyediakan informasi terkait distribusi topik penelitian mahasiswa sebagai bahan referensi untuk melihat arah penelitian yang sering atau jarang dilakukan, serta sebagai masukan dalam penyusunan *roadmap* penelitian.
 3. Bagi peneliti selanjutnya: Menjadi dasar untuk pengembangan penelitian lanjutan, baik dalam analisis tren topik, pengembangan metode *topic modeling*, maupun pengembangan sistem rekomendasi topik penelitian

1.4 Batasan Masalah

Agar penelitian ini terarah dan tidak menyimpang dari tujuan yang telah ditetapkan, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan berupa abstrak skripsi Program Studi Teknik Informatika dari 10 perguruan tinggi di Indonesia yang dapat diakses secara publik, dengan total 14.629 abstrak periode 2013–2025. Disajikan pada Tabel 1.2.

Tabel 1. 1 Daftar perguruan tinggi

No	Nama Perguruan Tinggi	Jumlah Data
1	Universitas Brawijaya (UB)	4343
2	Universitas Teknologi Digital Indonesia (UTDI)	2436
3	Universitas Komputer Indonesia (UNIKOM)	1945
4	Institut Teknologi Bandung (ITB)	1543
5	Universitas Atma Jaya Yogyakarta (UAJY)	1464
6	Institut Teknologi Nasional (ITN)	996
7	Universitas Surabaya (UBAYA)	708
8	Universitas Negeri Malang (UM)	534
9	Universitas Jenderal Soedirman (UNSOED)	421
10	Universitas Pelita Harapan (UPH)	238

2. Atribut data yang digunakan terbatas pada abstrak, tahun publikasi, dan asal perguruan tinggi. Atribut lain seperti judul, nama mahasiswa, atau kata kunci tidak dianalisis secara utama.
3. Metode pemodelan topik yang digunakan adalah BERTopic dengan *transformer-based sentence embedding* (model *paraphrase-multilingual-MiniLM*) tanpa melakukan perbandingan mendalam dengan seluruh metode *topic modeling* lainnya.
4. Dataset bersifat statis. Penelitian ini tidak mencakup pembaruan model secara inkremental apabila terdapat penambahan data baru setelah proses pemodelan selesai dilakukan.
5. Analisis tren dilakukan berdasarkan distribusi topik terhadap dimensi waktu (tahun publikasi) dan tidak mencakup analisis prediksi tren di masa mendatang (*forecasting*)

6. Definisi Topik: Dalam penelitian ini, "topik" didefinisikan secara operasional sebagai klaster semantik yang dihasilkan oleh BERTopic, yang merepresentasikan kumpulan dokumen dengan pola kemunculan istilah serupa. Topik bukan merupakan representasi formal dari bidang minat, rumpun keilmuan, atau konsentrasi akademik tertentu.
7. Interpretasi Topik: Proses pelabelan dan interpretasi topik dilakukan secara manual oleh peneliti berdasarkan kata kunci dominan hasil c-TF-IDF dan contoh dokumen representatif. Interpretasi ini bersifat deskriptif dan tidak dimaksudkan sebagai klasifikasi formal ke dalam taksonomi keilmuan nasional.
8. Evaluasi kualitas topik tidak menggunakan ambang batas absolut (*threshold*) untuk Cv dan TD. Pemilihan model terbaik dilakukan berdasarkan perbandingan nilai Cv dan TD antar skenario eksperimen.
9. Sistem yang dibangun berfungsi sebagai prototipe visualisasi dan eksplorasi hasil analisis, bukan sebagai sistem rekomendasi atau sistem prediksi otomatis.

1.5 Sistematika Penulisan

Adapun sistematika penulisan proposal yang digunakan terdiri dari beberapa bagian sebagai berikut:

1. Bab I PENDAHULUAN

Bab ini berisi latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, serta sistematika penulisan.

2. Bab II KAJIAN PUSTAKA

Bab ini memuat kajian terhadap penelitian terdahulu yang relevan, serta teori-teori yang mendukung penelitian, meliputi konsep *text mining*, *Natural Language Processing* (NLP), *topic modeling*, *transformer-based embedding*, BERTopic, serta analisis tren dan evolusi topik.

3. Bab III METODE PENELITIAN

Bab ini menjelaskan tahapan penelitian yang dilakukan, meliputi pengumpulan data abstrak skripsi, proses *preprocessing*,

pembentukan *embedding*, pemodelan topik menggunakan BERTopic, analisis tren berdasarkan tahun, serta perancangan prototipe visualisasi hasil analisis.

4. Bab IV HASIL DAN PEMBAHASAN

Bab IV menyajikan hasil penelitian yang diperoleh dari proses implementasi dan pengujian yang telah dilakukan. Pada bab ini ditampilkan hasil eksperimen, performa sistem atau model berdasarkan metrik evaluasi yang digunakan, serta perbandingan dengan metode pembandingan atau penelitian sebelumnya. Selain itu, bab ini juga memuat pembahasan dan analisis terhadap hasil yang diperoleh untuk menjelaskan keterkaitannya dengan tujuan dan rumusan masalah penelitian.

5. Bab V KESIMPULAN DAN SARAN

Bab V berisi kesimpulan akhir penelitian yang menjawab rumusan masalah dan tujuan penelitian berdasarkan hasil dan pembahasan pada bab sebelumnya. Selain itu, bab ini juga memuat saran untuk pengembangan penelitian selanjutnya, baik dari sisi metode, data, maupun sistem, agar penelitian dapat ditingkatkan atau dikembangkan lebih lanjut