BAB I

PENDAHULUAN

1.1 Latar Belakang

Knowledge Discovery in Databases (KDD) muncul dari kebutuhan untuk menganalisis data dalam jumlah besar (Shu dan Ye, 2023). KDD merupakan proses memperoleh pengetahuan dari data dengan menekankan pada penerapan metode-metode *data mining* (Jansevskis dan Osis, 2023). KDD juga mengacu pada proses *non-trivial* untuk mengidentifikasi pola-pola yang valid, baru, berpotensi bermanfaat, dan yang pada akhirnya dapat dipahami, dari sekumpulan data (Fayyad *et al.*, 1996, sebagaimana dikutip dalam Shu dan Ya, 2023). Proses KDD melibatkan lima tahapan utama, mulai dari seleksi dan prapemrosesan data, transformasi data, penerapan algoritma *data mining*, hingga interpretasi/evaluasi hasil. Pada tahapan *data mining*, berbagai algoritma *machine learning* digunakan untuk menemukan hubungan dalam data (Głowania *et al.*, 2023; Palacios *et al.*, 2021). Namun, tidak semua data dalam *data mining* bersifat terstruktur. Selain data numerik yang terstruktur, *data mining* juga menyediakan alat, metode, dan model untuk menganalisis data tidak terstruktur.

Text Mining merupakan metode yang populer dalam analisis data yang bersifat tidak terstruktur, khususnya data teks (Shu dan Ye, 2023). *Text Mining*, yang juga dikenal sebagai *Knowledge Discovery in Text Database*, merujuk pada proses menemukan pola dan pengetahuan tersembunyi yang menarik dan bermanfaat dari data teks dengan jumlah besar (Chai *et al.*, 2023). Secara umum, tahapannya serupa dengan KDD, namun karena teks merupakan data tidak terstruktur, penanganannya menjadi lebih kompleks dan memerlukan penerapan teknik-teknik *Natural Language Processing* (NLP), seperti *text preprocessing* untuk membersihkan dan meningkatkan kualitas data yang dilakukan dengan teknik-teknik seperti *case folding*, *filtering*,

tokenization, *stop words removal*, dan *lemmatization* (Siino *et al.*, 2024). Setelah itu dilakukan pula *feature extraction* untuk transformasi teks ke representasi numerik seperti *Bag-of-Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) agar data tersebut dapat dipahami oleh mesin (Aleqabie *et al.*, 2024; Birunda dan Devi, 2021; Khairunnisa *et al.*, 2021).

Salah satu teknik penting dalam *text mining* adalah *topic modeling*, yaitu metode *clustering* yang digunakan untuk mengidentifikasi struktur topik dalam koleksi teks dengan jumlah besar (de Lima *et al.*, 2023). *Topic modeling* salah satunya diaplikasikan untuk analisis tren penelitian di berbagai bidang dengan menganalisis kumpulan dokumen penelitian dalam jumlah besar (Takacs dan O'Brien, 2023). Dua algoritma *topic modeling* yang banyak digunakan adalah *Latent Dirichlet Allocation* (LDA) dan BERTopic (de Lima *et al.*, 2023; Niroomand *et al.*, 2023).

LDA merupakan metode statistik yang digunakan untuk *topic modeling*, dengan tujuan mengidentifikasi pola-pola yang berulang dalam koleksi dokumen (de Lima *et al.*, 2023). LDA menganggap setiap dokumen terdiri dari berbagai topik, dan setiap topik terdiri dari berbagai kata. Dengan menggunakan kata dan dokumen sebagai input, LDA belajar untuk mengidentifikasi topik-topik melalui *text mining* tanpa pengawasan (*unsupervised*) (Shu dan Ye, 2023). Meskipun populer, LDA dianggap memiliki kelemahan karena mengabaikan hubungan semantik antar kata dan tidak memperhitungkan konteks kata dalam kalimat, sehingga dapat menyebabkan kumpulan kata pada data input tidak dapat mewakili suatu dokumen secara akurat (H. Lee *et al.*, 2023).

BERTopic merupakan model *topic modeling* berbasis *embedding* yang dibangun di atas BERT, yang dapat menghasilkan representasi vektor kata dan kalimat dengan properti semantik, sehingga memungkinkan pemahaman secara kontekstual (Kukushkin *et al.*, 2022). BERTopic bekerja dengan membuat *embedding* dokumen, kemudian mereduksi dimensi data dan mengelompokkan data (*clustering*), dilanjutkan dengan ekstraksi topik dengan mempertimbangkan relevansi kata dalam setiap kelompok (*cluster*) (de Lima *et al.*, 2023; Niroomand *et al.*, 2023). BERTopic mampu mempertahankan properti semantik dokumen dalam topik, sehingga dapat menghasilkan topik yang lebih beragam dan koheren dibandingkan LDA, serta lebih *robust* dalam penggunaan, tidak terlalu bergantung dengan *preprocessing*, dan memungkinkan untuk dilakukannya *fine-tuning* (de Lima *et al.*, 2023).

Algoritma *topic modeling* seperti LDA dan BERTopic telah banyak digunakan dalam menganalisis berbagai data tekstual, seperti data *tweet* (de Lima *et al.*, 2023; Ng *et al.*, 2023; Ogunleye *et al.*, 2023; Y. Wang *et al.*, 2023), *direct message* di media sosial (Khadija dan Nurharjadmo, 2024), teks berita (Chen *et al.*, 2023; H. Lee *et al.*, 2023), dan buku teks (Karadağ, 2023). Kedua metode ini juga sudah diterapkan pada data publikasi penelitian di berbagai bidang untuk menganalisis topik penelitian dan memahami trennya dari waktu ke waktu. LDA telah digunakan untuk meneliti tren di bidang *sustainability* dan *marketing* (Jung dan Kim, 2023), penerapan fuzzy (Yu *et al.*, 2023), serta layanan ekosistem dan biodiversitas (Takacs dan O'Brien, 2023). Sementara itu, BERTopic telah digunakan dalam penelitian sebelumnya untuk menganalisis tren di bidang NLP (Samsir *et al.*, 2023), penerapan *Artificial Intelligence* (AI) dalam sistem energi terbarukan (Niroomand *et al.*, 2023), dan AI *maturity* (Akbarighatar *et al.*, 2023). Namun, penerapan kedua metode ini untuk menganalisis tren penelitian dalam bidang ilmu yang lebih luas seperti Ilmu Komputer masih terbatas, terlepas dari popularitasnya.

Penelitian di bidang Ilmu Komputer menunjukkan kemajuan yang pesat dan terus berkembang secara dinamis dalam beberapa dekade ini (Qin *et al.*, 2021). Khususnya dalam kurun waktu lima tahun terakhir (2019-2023), telah tercatat lebih dari 2,9 juta publikasi penelitian di bidang Ilmu Komputer, menempatkannya pada peringkat ketiga dengan kontribusi sebesar 8,2% dari total publikasi penelitian di seluruh dunia, setelah bidang Kedokteran (14%) dan Teknik (11,5%). Bahkan, dalam tiga tahun terakhir (2021-2023), persentase penelitian dalam Ilmu Komputer meningkat signifikan hingga mencapai 10,75% peningkatan rata-rata pertahunnya (Scopus, n.d.). Meskipun pertumbuhannya pesat, analisis tren penelitian dalam bidang Ilmu Komputer ternyata masih terbilang kurang. Hal ini menyebabkan para peneliti kesulitan untuk memahami dengan jelas arah perkembangan dan prospek topik penelitian di masa mendatang (Qin *et al.*, 2021). Kajian mendalam terhadap berbagai topik penelitian menjadi sangat penting untuk memahami tahapan pembentukan pengetahuan dan evolusi ilmu, sekaligus menilai kualitas suatu disiplin ilmu serta dampaknya di dunia akademis (Widyaningsih *et al.*, 2021). Oleh karena itu, kebutuhan akan analisis tren dan identifikasi topik penelitian menjadi isu menarik yang perlu dibahas dalam penelitian ini.

Penelitian sebelumnya yang mengkaji bidang ini masih mengandalkan metode konvensional seperti analisis bibliometrik dan pengelompokan topik berdasarkan

subbidang penelitian (Qin *et al.*, 2021; Widyaningsih *et al.*, 2021). Padahal, analisis tren topik penelitian yang lebih komprehensif dengan menggunakan metode *topic modeling* seperti LDA dan BERTopic sangat penting untuk memperoleh pemahaman yang mendalam tentang tren topik penelitian di bidang Ilmu Komputer. Hal ini penting karena dapat mendukung perencanaan penelitian yang efisien di masa depan (Widyaningsih *et al.*, 2021; Yukselen *et al.*, 2022).

Dengan demikian, masalah yang akan diteliti dalam penelitian ini adalah bagaimana menerapkan metode *topic modeling*, khususnya menggunakan BERTopic dan LDA, untuk mengidentifikasi topik-topik dalam penelitian di bidang Ilmu Komputer secara komprehensif. Penelitian ini juga akan mengevaluasi efektivitas kedua metode tersebut dalam mengidentifikasi topik-topik penelitian di bidang Ilmu Komputer, serta menganalisis dinamika perkembangan tren topik-topik penelitian dari tahun ke tahun. Algoritma BERTopic dipilih karena memanfaatkan model bahasa BERT untuk pemahaman kontekstual dan mempertahankan hubungan semantik antar kata, menghasilkan topik-topik yang lebih akurat (de Lima *et al.*, 2023; H. Lee *et al.*, 2023). Meskipun demikian, LDA tetap dipilih sebagai metode pembanding karena keandalannya yang sudah teruji dalam pemodelan topik (Shu dan Ye, 2023), sehingga memungkinkan evaluasi terhadap kelebihan dan kekurangan kedua metode ini dalam mengidentifikasi topik-topik penelitian di bidang Ilmu Komputer.

Berdasarkan uraian latar belakang di atas, maka penelitian ini diangkat dan dituangkan ke dalam bentuk tesis yang berjudul "**Analisis Tren Penelitian Bidang Ilmu Komputer dengan Metode BERTopic dan LDA**".

1.2 Perumusan Masalah

Berdasarkan latar belakang permasalahan yang ada, maka rumusan masalah yang dibahas pada penelitian ini adalah:

1. Bagaimana menerapkan metode *topic modeling*, khususnya menggunakan BERTopic dan LDA, untuk mengidentifikasi topik-topik dalam penelitian di bidang Ilmu Komputer?
2. Bagaimana efektivitas metode *topic modeling* dengan BERTopic dan LDA, dalam mengidentifikasi topik-topik penelitian di bidang Ilmu Komputer?

3. Bagaimana dinamika perkembangan tren topik-topik penelitian di bidang Ilmu Komputer dari tahun 2019 hingga 2023?

1.3 Batasan Masalah

Batasan masalah mencakup ruang lingkup serta keterbatasan yang ada dalam proses penelitian ini, diuraikan sebagai berikut:

1. Data yang digunakan dalam penelitian ini bersumber dari situs Emerald Insight, mencakup metadata artikel penelitian di bidang Ilmu Komputer yang diterbitkan dalam periode lima tahun terakhir, yaitu dari tahun 2019 hingga 2023.
2. Pengumpulan data dilakukan dengan metode pencarian berbasis *query* menggunakan 118 kata kunci (*keyword*) dalam Bahasa Inggris yang terkait dengan Ilmu Komputer. Kata kunci ini dibagi ke dalam 12 sesi pencarian. Setiap sesi pencarian menghasilkan 500 data, sehingga total data yang terkumpul berjumlah 6.000 data.
3. Bahasa artikel penelitian yang dipilih terbatas pada Bahasa Inggris yang merupakan bahasa utama pada jurnal-jurnal global terakreditasi.
4. Fokus analisis topik terbatas pada data teks judul dan abstrak artikel penelitian sebagai representasi inti dari informasi pada artikel penelitian.
5. Penelitian hanya memanfaatkan bahasa pemrograman Python untuk analisis data dan implementasi metode *topic modeling*, khususnya BERTopic dan LDA, menggunakan platform Google Colaboratory.
6. Hasil penelitian berupa model topik yang dapat digunakan untuk mengidentifikasi topik penelitian di bidang Ilmu Komputer serta analisis trennya dalam bentuk visualisasi. Kedua hasil ini diimplementasikan dalam bentuk aplikasi sederhana berbasis web.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mencapai tiga sasaran utama yang telah dirumuskan, dengan perincian sebagai berikut:

1. Menerapkan metode *topic modeling*, khususnya BERTopic dan LDA, untuk mengidentifikasi topik-topik yang muncul dalam penelitian di bidang Ilmu Komputer.
2. Mengevaluasi efektivitas penerapan metode *topic modeling* dengan BERTopic dan LDA dalam mengidentifikasi topik-topik penelitian di bidang Ilmu Komputer.
3. Menganalisis dinamika perkembangan tren topik-topik penelitian di bidang Ilmu Komputer dari tahun 2019 hingga 2023.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat yang signifikan bagi berbagai pihak, baik bagi peneliti, objek penelitian, maupun bagi penelitian selanjutnya. Berikut adalah uraian manfaat dari penelitian ini:

1. Bagi Peneliti:

   Penelitian ini diharapkan dapat meningkatkan pemahaman dan keterampilan peneliti dalam menerapkan metode *topic modeling*, khususnya BERTopic dan LDA, untuk analisis tren di bidang Ilmu Komputer. Dengan memanfaatkan kedua metode ini, peneliti akan memperoleh wawasan yang lebih mendalam mengenai teknik pemodelan topik yang efektif dalam mengidentifikasi dan menganalisis topik-topik penelitian yang relevan.
2. Bagi Objek Penelitian (Bidang Ilmu Komputer):

   Penelitian ini memberikan kontribusi pada pemetaan dan pemahaman yang lebih komprehensif terhadap perkembangan tren topik penelitian di bidang Ilmu Komputer selama periode 2019-2023. Identifikasi tren topik yang naik dan turun dapat membantu komunitas akademisi dan industri dalam mengarahkan fokus penelitian serta pengembangan di masa depan.

3. Bagi Penelitian Selanjutnya:

   Penelitian ini memberikan landasan yang kuat bagi studi-studi di masa depan dalam mengembangkan dan menyempurnakan metode *topic modeling* untuk analisis tren penelitian. Dengan membuktikan efektivitas BERTopic dan LDA, penelitian ini dapat menjadi acuan bagi penelitian selanjutnya untuk memperluas aplikasi metode ini di berbagai disiplin ilmu dan periode waktu lainnya yang lebih panjang.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam dokumen penelitian ini akan dijelaskan dalam susunan berikut:

BAB I PENDAHULUAN

Pada bab ini diuraikan latar belakang, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan pada penelitian ini.

BAB II LANDASAN TEORI

Bab ini memaparkan kerangka teoritis atau konsep-konsep dasar yang menjadi landasan penelitian serta daftar beberapa penelitian terdahulu yang berkaitan dengan penelitian ini, dalam bentuk *state-of-the-art* penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai kerangka penelitian dan langkah-langkah yang dilakukan dalam proses penelitian. Mulai dari identifikasi dan analisis masalah, studi literatur, pengumpulan data, analisis data, perancangan aplikasi, implementasi dengan *tool*, pengujian, hingga hasil dan pembahasan.

BAB IV ANALISA DAN PERANCANGAN

Pada bab ini, dipaparkan tahapan analisis data, termasuk analisis algoritma yang akan diimplementasikan, serta cuplikan hasil analisis

dan uraian pembahasannya. Pada bab ini juga ditunjukkan tahapan perancangan aplikasi dan hasilnya berupa desain tampilan dan *flowchart* aplikasi.

BAB V IMPLEMENTASI DAN HASIL

Bab ini menguraikan tahapan implementasi algoritma yang telah dijelaskan pada Bab IV, serta hasil yang diperoleh. Pembahasan berfokus pada hasil implementasi akhir, yaitu *output* penelitian berupa analisis tren dan aplikasi prediksi topik.

BAB VI KESIMPULAN DAN SARAN

Bab terakhir ini akan merangkum kesimpulan dari seluruh proses penelitian yang telah dilakukan. Kesimpulan akan memuat jawaban atas rumusan masalah penelitian, temuan-temuan utama, dan implikasi dari hasil penelitian. Selain itu, bab ini juga akan menyajikan saran-saran untuk penelitian selanjutnya atau untuk pengembangan lebih lanjut terkait topik yang diteliti.