BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era informasi digital, ketersediaan sumber konten digital, khususnya dalam bentuk publikasi ilmiah mengalami peningkatan jumlah secara eksponensial. Menurut sebuah penelitian, tingkat pertumbuhan publikasi ilmiah yang diambil dari empat sumber (*Dimensions*, *Microsoft Academic*, *Scopus*, dan *Web of Science*) mencapai 4,10% dengan waktu penggandaan 17,3 tahun (Bornmann el al., 2021). Pertumbuhan eksponensial pada publikasi ilmiah memiliki efek yang signifikan pada komunitas sains, khususnya pada area penelitian dan evaluasi ilmiah. Salah satu cara untuk mengakses literatur ilmiah adalah melalui repositori atau *database* penyimpanan online yang merupakan platform menyediakan akses ke literatur ilmiah, termasuk makalah penelitian, artikel dan komunikasi ilmiah lainnya.

Menurut sebuah penelitian, peningkatan dalam jumlah publikasi ilmiah yang relevan membuat hampir tidak mungkin untuk seorang peneliti untuk mengikutinya bahkan jika mereka meluangkan sebagian besar waktu produktifnya (Kraus et al., 2023). Pada pertumbuhan penelitian terhadap bisnis kecil dan kewirausahaan, peneliti menghadapi tantangan besar dalam mempertahankan gambaran yang komprehensif terhadap perkembangan terbaru di bidang tertentu (Kraus et al., 2023).

ArXiv adalah repositori sumber terbuka yang sangat luas dengan akses ke ribuan publikasi ilmiah dari berbagai disiplin ilmu. Dalam penelitian ini, penulis memilih ilmu komputer karena representasi kuatnya dalam perkembangan teknologi.

Pemodelan topik merupakan teknik *machine learning* yang bertujuan untuk mengidentifikasi dan menganalisa tema-tema atau topik-topik yang mendasar pada suatu kumpulan dokumen. Pemodelan topik melibatkan *clustering* atau pengelompokan kata-kata yang cenderung sering muncul secara bersamaan pada suatu kumpulan dokumen. Algoritma *clustering* menawarkan pendekatan yang bertujuan untuk menyelesaikan masalah ini dengan mengelompokkan publikasi ilmiah dengan topik yang serupa sehingga mempermudah peneliti untuk menganalisa studi terkait secara lebih efisien (Thrun & Stier, 2021).

Salah satu teknik pemodelan topik yang sering digunakan adalah LDA (*Latent Dirichlet Allocation*). LDA adalah metode statistik yang digunakan untuk mengidentifikasi topik-topik yang tersembunyi pada suatu kumpulan dokumen teks. LDA merupakan model probabilistik generatif unutuk pengumpulan data diskrit sebagai sebuah korpus atau kumpulan kata (Blei el al., 2003). Teknik pemodelan topik lainnya adalah NMF (*Non-Negative Matrix Factorization*), NMF merupakan pendekatan *unsupervised* yang efektif dalam menemukan topik mendasar dalam suatu kumpulan teks (Alfajri dkk, 2022). Akan tetapi, salah satu batasan dari LDA dan NMF adalah ketergantungan pada reprentasi kumpulan kata yang mengabaikan hubungan semantik antar kata. Karena representasi ini mengabaikan hubungan antara kata-kata dalam suatu kalimat (Grootendorst, 2022).

Untuk mengatasi masalah ini, teknik *text embedding* menjadi populer di ranah NLP (*Natural Languange Processing*), untuk lebih spesifiknya, BERT (*Bidirectional Encoder Representations from Transformer*) (Devlin et al., 2019) dan banyak variasinya, menunjukkan hasil yang sangat baik dalam menghasilkan kalimat dan representasi vektor kontekstual (Grootendorst, 2022).

Salah satu variasi BERT adalah *BERTopic* yang merupakan salah satu *pre-trained unsupervised language models* yang memanfaatkan BERT *embeddings* dan menggunakan c-TF-IDF (*Class-Based Term Frequency-Inverse Document Frequency*) untuk membuat klaster yang padat. *BERTopic* yang dikembangkan oleh Grootendorst (2022), menghasilkan topik melalui tiga langkah. Mengubah tiap dokumen menjadi

representasi *embedding* kemudian sebelum melakukan *clustering* pada *embedding* tersebut, dimensionalitasnya dikurangi untuk mengoptimasi proses *clustering* dan yang terakhir dari klaster dokumen, representasi topik diekstrak menggunakan variasi berbasis kelas kustom dari TF-IDF (Grootendorst, 2022).

Salah satu penelitian yang menerapkan *BERTopic* adalah oleh Samsir dkk (2023), yang berhasil melakukan penerapan topic modellling pada abstrak penelitian di bidang NLP. Secara keseluruhan, penelitian tersebut menunjukkan efektifitas *BERTopic* dalam mengekstrak makna dari kumpulan korpus abstrak akademik yang besar.

Mengetahui hal tersebut, penulis mengajukan proposal tugas akhir dengan judul "IMPLEMENTASI *BERTOPIC* PADA ABSTRAK PUBLIKASI ILMIAH BIDANG ILMU KOMPUTER", untuk menggali topik-topik dan tema dari berbagai disiplin ilmu yang berbeda dari abstrak publikasi ilmiah yang terbuka secara umum menggunakan *BERTopic*.

1.2 Rumusan Masalah

Dengan besarnya volume literatur ilmiah yang ada sekarang, membuat tantangan tersendiri untuk menemukan topik yang relevan:

1. Eksplorasi topik secara manual dengan membaca abstrak memakan waktu yang banyak. Apakah BERTopic dapat mengidentifikasi dan mengekstrak topik yang relevan dari kumpulan publikasi ilmiah melalui abstrak?
2. Apakah BERTopic dapat diterapkan untuk menemukan tren?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

1. Mengidentifikasi topik-topik utama yang muncul berdasarkan kumpulan abstrak menggunakan BERTopic.
2. Mengidentifikasi tren penelitian di penelitian ilmu komputer di tahun 2023.

1.3.2 Manfaat

1. Membantu dalam mengidentifikasi topik yang relevan dan melihat tren pada disiplin ilmu yang berbeda.
2. Menambah wawasan terkait pemodelan topik menggunakan teknik BERTopic pada abstak literatur ilmiah di bidang ilmu komputer.

1.4 Batasan Masalah

1. Dataset yang digunakan diambil dari repositori arXiv dengan rentang waktu pengambilan per Januari sampai dengan Desember 2023 sesuai dengan tanggal submit paper yang tertera di arXiv.
2. Data yang digunakan berupa abstrak publikasi ilmiah bidang ilmu komputer.

1.5 Teori

1.5.1 *Natural Language Processing*

Bahasa adalah sebuah sarana untuk menyampaikan pemikiran, informasi, dan ide bersamaan dengan perasaan, ketidaksempurnaan, dan ambiguitas (Johri dkk., 2021). Karenanya, sulit untuk memandang bahasa sebagai sesuatu yang memiliki kombinasi

peraturan matematis. NLP (*Natural Language Processing*) adalah area penelitian yang befokus pada bagaimana komputer dapat digunakan untuk memahami dan memanipulasi bahasa baik secara tekstual atau secara lisan. Tujuan utama dari NLP adalah untuk mempelajari bagaimana manusia memahami dan menggunakan bahasa agar alat dan teknik yang dapat membuat komuter memahami dan memanipulasi bahasa untuk menyelesaikan suatu pekerjaan dapat dikembangkan (Chowdhary, K. R, 2020).

Penggunaan NLP pertama kali dilakukan oleh bangsa Jerman pada Perang Dunia II. Sebuah mesin bernama enigma digunakan untuk menyamarkan pesan rahasia Jerman menjadi kode rahasia untuk kemudian disampaikan ke pasukan militer Jerman yang ditempatkan di Eropa (Johri dkk., 2021). Kemudian muncullah algoritma seperti *decision tree* yang menggunakan aturan *if-then*. Kini, NLP mengarah ke *deep learning*. Disebabkan oleh kemampuan *deep learning* dalam menyelesaikan tugas yang sulit dilakukan dengan hanya menggunakan aturan yang tetap. NLP memiliki dapat diterapkan di banyak bidang. Dengan adanya *deep learning* teknologi seperti Cortana Alexa, Siri, dan Google Assistant sudah dapat dilihat di sekitar kita baik di rumah ataupun di lingkungan kerja. NLP juga digunakan untuk menganalisa data tekstual dalam jumlah yang besar secara cepat dan efisien (Johri dkk., 2021).

1.5.2 *Text Mining* dan *Web Scraping*

Kehidupan sekarang sangat erat kaitannya dengan dunia digital. Kegiatan kecil pun dapat menghasilkan data digital, seperti berbelanja tiket secara online, data tersebut akan tersimpan ke dalam database. Diperkirakan 80% dari data digital berupa data teks. Data ini terdiri dari data tidak terstuktur, data tidak berguna, data berguna, data ilmiah, dan lain-lain. Mengolah data teks kemudian menemukan informasi penting didalamnya biasanya dikenal dengan *Text mining* (Kaushik & Naithani, 2016). *Text mining* dan *data mining* itu hal yang sama, akan tetapi *data mining* bekerja pada data yang terstruktur, di sisi lain *text mining* bekerja pada data semi-terstruktur dan tidak terstruktur (Ajay Jadhav et al., 2023).

Ada beberapa teknik dalam area bahasan *text mining*, beberapa diantaranya adalah *summarization,* ekstraksi informasi, kategorisasi teks, *clustering* dan lainnya. Pengaplikasiannya juga beragam, seperti analisis sentimen, analisis tren dan *fraud detection*. Tiap pengaplikasian tersebut cenderung akan menghadapi masalah yang serupa yaitu mengolah data yang tidak terstruktur dan berurusan dengan data dengan voluma yang sangat besar.

Abstrak dari publikasi ilmiah merupakan bagian sangat penting. Banyak informasi yang dapat didapatkan dalam membaca abstrak publikasi ilmiah, seperti masalah yang dihadapi, metode yang diusulkan untuk menyelesaikan masalah, penggunaan datan, dan hasil dari penelitian. Salah satu metode yang dapat digunakan untuk memperoleh data abstrak publikasi ilmiah adalah *web scraping* yaitu dengan mengambil informasi dari suatu web (Zhao, 2017). Dengan melakukan *web scraping*, informasi-informasi dapat diperoleh yang kemudian akan diolah untuk mendapatkan informasi yang berharga.

1.5.3 *Transformer*

Transformer adalah jenis *deep learning* yang diperkenalkan pertama kali pada tahun 2017 oleh Vaswani dkk. dalam paper berjudul *“Attention is All You Need”.* Model *transformer* disebutkan dapat menimbang pentingnya kata-kata yang berbeda dalam suatu kalimat dengan menggunakan konsep *self-attention*. Konsep ini membuat *transformer* dapat memperoleh hubungan semantik antar kata secara lebih efektif

dibandingkan dengan model tradisional (Wolf et al., 2020). Model ini menggunakan kalimat sebagai masukannya kemudian diubah menjadi *token* kemudian tiap *token* di konversi menjadi vektor berdimensi tinggi untuk menghasilkan makna kontekstual dari kumpulan kata. Produk dot dari *self-attention* adalah *attention score* dengan setiap token dalam urutan. *Transformer* terdiri dari dua bagian yaitu *encoder* yang memproses urutan masukan dan menghasilkan representasi dari tiap token dan *decoder* yang menghasilkan urutan luaran dengan memerhatikan hasil luaran *encoder* dan token yang dihasilkan sebelumnya (Acheampong et al., 2021).

Beberapa model yang dikembangkan berdasarkan *transformer* adalah model *state-of-the-art* seperti BERT (*Bidirectional Encoder Representations from Transformers*), yang menggunakan bagian *encoder* dari *transformer*, dirancang untuk memahami konteks teks secara mendalam dan melakukan analisis teks. Sebaliknya GPT (*Generative Pre-Training*) yang menggunakan bagian *decoder* dari *transformer*, berfokus pada generasi teks. Sementara itu, T5 (*Text-To-Text Transfer Transformer*) yang menggunakan keduanya, dirancang untuk melakukan tugas NLP dalam format *text-to-text*, seperti terjemahan text, rangkuman hingga pelabelan dengan mengubah format input dan output.

BERT dengan kekuatannya menghasilkan *text embeddings* yang memahami hubungan semantik antar teks, lebih cocok digunakan dalam aplikasi yang berfokus pada representasi teks dibandingkan T5 yang berfokus pada transformasi teks. *Transformer* memiliki fleksibilitas dalam melakukan tugas yang kompleks terkait dengan NLP seperti mesin penerjemah dengan menerjemahkan kalimat antar bahasa dengan mempertahankan maknanya dan pada *text summarization* dengan mengerucutkan teks dengan volume besar menjadi sebuah kesimpulan dengan mempertahankan informasi penting.

1.5.4 BERT

BERT (*Bidirectional Encoder Representations from Transformer*) adalah model berbasis *deep learning* yang bertujuan untuk menganalisa teks secara *bidirectional* atau dua arah. Alih-alih membaca teks dalam satu arah atau dari kiri ke kanan yang membuat model hanya memiliki informasi tentang kata yang ada sebelumnya dan tidak dengan kata yang ada setelahnya, BERT menggunakan arsitektur *transformer* untuk memproses teks secara dua arah. Dengan *transformer* BERT menjadi populer dan menarik banyak perhatian peneliti di bidang NLP untuk diterapkan pada beberapa masalah untuk mendapatkan hasil *state of the art* dari metode lainnya (Aftan dan Shah, 2023).

BERT merupakan *pre-trained model* yang dilatih menggunakan dua tugas utama yaitu MLM (*Masked Language Modelling*) dimana BERT ditugaskan untuk memprediksi kata acak yang disamarkan dalam satu kalimat dan NSP (*Next Sentence Prediction*) dimana BERT dilatih untuk memahami hubungan antar kalimat dengan memprediksi jika satu kalimat mengikuti kalimat lain dalam suatu teks, ini membantu dalam mendapatkan pemahaman antara makna lintas kalimat. Arsitektur model BERT berdasar pada *multi-layer bidirectional transformer encoder*. Tiap lapisan terdiri dari beberapa *self-attention heads*, dan tiap lapisan ini berfungsi untuk memahami lebih baik suatu kalimat ibarat membaca kalimat itu berulang kali dan lebih memahaminya sedikit demi sedikit. Mekanisme ini membantu BERT dalam memahami hubungan antara kalimat dan memanfaatkan konteks dari dua arah (Devlin et al., 2019).

1.5.5 Pemodelan Topik

Pemodelan topik adalah metode analisa yang populer digunakan untuk mengevaluasi data berbasis teks dan menemukan kumpulan kata-kata yang relevan (dimasukkan ke dalam kumpulan bernama topik) dari koleksi dokumen. Analisa ini dapat membantu dalam mengolah data berskala besar seperti mengola data jurnal atau artikel dengan fokus yang serupa, atau analisa pengguna sosial media dengan postingan yang serupa. Tetapi di sisi lain, pemodelan topik kerap berhadapan dengan masalah seperti optimasi dan *noise sensitivity*. Oleh karena itu, banyak motode yang telah dikembangkan untuk pemodelan topik dengan pertimbangan ukuran dan variasi kumpulan data yang akan diolah (Vayansky dan Kumar, 2020).

Metode yang umum digunakan adalah LDA (*Latent Dirichlet Allocation*) yang dikembangkan oleh Blei et al. pada tahun 2003. LDA merupakan metode yang bertujuan untuk menemukan pola pada suatu kumpulan dokumen untuk menghasilkan topik menggunakan metode statistik. Pengaplikasian LDA bervariasi mulai dari menemukan tema tersembunyi dari kumpulan dokumen, membuat kesimpulan dari teks dalam jumlah yang besar dan mengorganisir dokumen berdasarkan topik. Meskipun LDA layak digunakan untuk pemodelan topik menggunakan data yang bervariasi, LDA tidak mempu untuk memodelkan relasi data yang lebih lanjut dan memiliki performa yang buruk saat dokumen tidak memiliki panjang yang sesuai (Vayansky dan Kumar, 2020).

BERTopic adalah salah satu variasi dari BERT (*Bidirectional Encoder Representations from Transformer*) yang merupakan salah satu *pre-trained unsupervised language models* yang memanfaatkan BERT *embeddings* dan c-TF-IDF (*Class-Based Term Frequency-Inverse Document Frequency*) untuk mengekstrak representasi topik yang koheren. Metode tradisional seperti LDA dan NMF dapat menemukan topik dalam suatu kumpulan teks, akan tetapi salah satu batasan dari metode ini adalah ketergantungannya pada representasi kumpulan kata yanag mengabaikan hubungan semantik antar kata dalam suatu kalimat (Grootendorst, 2022).

1.5.6 *BERTopic*

BERTopic merupakan teknik pemodelan topik yang memanfaatkan *transformers* dan c-TF-IDF (*Class-Based Term Frequency-Inverse Document Frequency*) untuk menghasilkan representasi topik. Dalam menggunakan *BERTopic*, dokumen disematkan dalam proses *embedding* untuk menghasilkan representasi data dalam ruang vektor. Biasanya data yang dihasilkan dari proses penyematan atau *embedding* ini setidaknya memiliki 384 dimensi (Grootendorst, 2022) dan banyak algoritma *clustering* memiliki kesulitan untuk memroses data dengan dimensi yang tinggi. Untuk itu, dilakukan pengurangan dimensi atau *dimensionality reduction* untuk merubah data vektor berdimensi tinggi menjadi berdimensi lebih rendah menggunakan UMAP (*Uniform Manifold and Projection*) (McInnes et al., 2020). UMAP bekerja dengan membuat grafik representasi data berdasarkan hubungan antar titik kemudian mengoptimasi grafik untuk menempatkan titik-titip serupa lebih dekat dalam ruang dimensi rendah sementara menempatkan titik-titik yang tidak serupa berjauhan.

Clustering atau klasterisasi adalah metode pengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum. Metode *clustering* yang digunakan dalam penelitian ini adalah HDBSCAN (*Hierarchical Density Based Spatial Clustering of Application with Noise*). HDBSCAN membuat klaster berdasarkan kepadatan distribusi data dan mengidentifikasi titik yang terisolasi atau memiliki

kepadatan rendah sebagai *noise* (Campello et al., 2013). Setelah proses *clustering BERTopic* menggunakan c-TF-IDF untuk mengekstrak representasi topik dari setiap *cluster*. c-TF-IDF merupakan variasi dari TF-IDF (*Term Frequency-Inverse Document Frequency*) yang umumnya digunakan untuk memodelkan pentingnya kata-kata dalam kelompok dokumen. Namun, dalam c-TF-IDF, pendekatan ini diperluas dengan menghitung pentingnya suatu kata dalam suatu *cluster* atau topik secara keseluruhan (Grootendorst, 2022).

1.5.7 Metrik Evaluasi

Topic coherence adalah metrik yang digunakan untuk menilai kualitas topik yang dihasilkan oleh algoritma pemodelan topik seperti LDA (*Latent Dirichlet Allocation*), NMF (*Non-negative Matrix Factorization*) dan *BERTopic*. *Topic coherence* mengukur seberapa baik kata-kata teratas atau *top words* dalam suatu topik saling terkait atau memiliki makna yang konsisten. Metrik ini menunjukkan seberapa relevan atau logis kata-kata dalam suatu topik jika dilihat secara bersamaan dengan tujuan untuk mencerminkan kualitas topik-topik yang dihasilkan dari perspektif interpretasi manusia (Blair et al., 2020). Hal ini penting karena metode pemodelan topik berbasis *unsupervised learning* tidak menjamin topik yang dihasilkan selalu koheren atau mudah diinterpretasikan (Röder et al, 2015). Metode perhitungan coherence menggunakan pendekatan berbasis probabilitas atau statistik seperti NPMI (*Normalized Pointwise Mutual Information*) sebagai dasar perhitungannya yang memiliki minimum 0 dan nilai maksimum 1 ketika kata-kata selalu muncul bersama-sama secara sempurna. Dengan menggunakan *topic coherence*, kita dapat mengevaluasi topik dari perspektif interpretasi manusia. Meskipun pengukuran *topic coherence* yang dibuat oleh manusia menjadi standar tertinggi dalam mengevaluasi model topik. Akan tetapi, pengukuran dengan cara ini mahal dan memakan waktu yang banyak (Röder et al, 2015). Röder et al. (2015) mengusulkan pendekatan untuk menghitung *coherence score* yang terdiri dari empat tahapan, yaitu:

1. *Segmentation*

Setiap kata teratas (*top words*) dari setiap topik, diambil dan dibagi ke dalam pasangan kata (*pairwise*). Misalnya, topik dengan daftar kata $[w_1, w_2, w_3, \dots\ w_n]$, pasangan kata yang terbentuk adalah $[(w_1, w_2), (w_1, w_3), \dots, (w_n, w_m)]$. Segmentasi ini bertujuan untuk menganalisis hubungan antar kata secara berpasangan.

2. *Probability Estimation*

Probabilitas dan hubungan antar kata dari hasil segmentasi dihitung berdasarkan frekuensi kemunculan kata-kata tersebut dalam *corpus*. Frekuensi ini dihitung menggunakan pendekatan *sliding window*, dimana jendela tertentu digunakan untuk menilai konteks kemunculan kata-kata tersebut.

3. *Measure Calculation*

Pada tahap ini, hubungan antar kata dalam pasangan dihitung menggunakan NPMI (*Normalized Pointwise Mutual Information*). NPMI digunakan untuk menilai seberapa sering pasangan kata muncul bersama dibandingkan dengan kemunculannya secara individual.

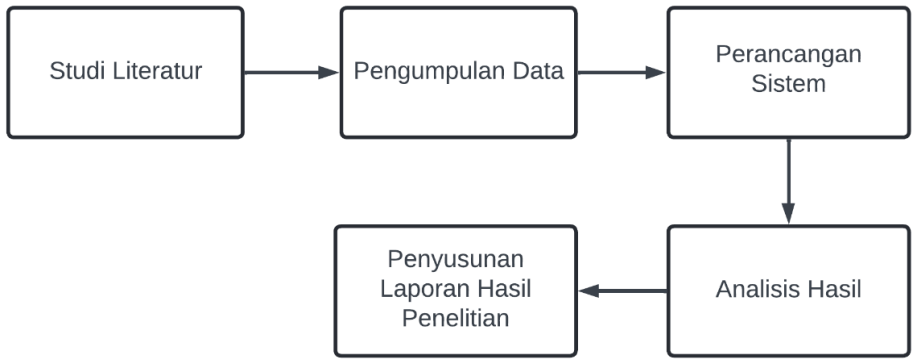
4. *Aggregation*

Coherence score untuk setiap pasangan kata digabungkan dengan cara dirata-ratakan untuk menghasilkan *coherence* keseluruhan dari topik tersebut.

BAB II
METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian “IMPLEMENTASI *BERTOPIC* PADA ABSTRAK PUBLIKASI ILMIAH BIDANG ILMU KOMPUTER” terdiri dari beberapa tahapan seperti pada Gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian

Garis besar tahapan penelitian dijelaskan sebagai berikut.

1. Studi Literatur

Pada tahap ini, peneliti melakukan kajian terhadap literatur yang relevan terkait *BERTopic* dan penerapannya dalam analisis teks ilmiah. Selain itu, peneliti mencari referensi terkait dengan metode pengolahan data dan teknik evaluasi model.

2. Pengumpulan Data

Data abstrak publikasi ilmiah yang digunakan dalam penelitian diperoleh dari repositori terbuka arXiv. Pengumpulan data dilakukan dengan teknik *scraping* pada kategori *Computer Science*. Data yang dikumpulkan terbatas pada publikasi tahun 2023.

3. Perancangan Sistem

Perancangan sistem untuk pemodelan topik menggunakan *BERTopic* mencakup beberapa proses utama seperti *preprocessing* untuk membersihkan dan mempersiapkan data yang diperoleh sebelum diolah, pemodelan dengan *BERTopic* dan *hyperparameter* tuning untuk mendapatkan hasil pemodelan yang optimal.

4. Analisis Hasil Penelitian

Setelah mendapatkan topik dari model *BERTopic*, dilakukan analisis terhadap topik-topik yang ditemukan. Analisis meliputi evaluasi kualitas topik menggunakan *coherence score*, visualisasi topik, dan evaluasi hasil pemetaan topik.

5. Penyusunan Laporan Hasil Penelitian

Tahap terakhir adalah penyusunan laporan terkait seluruh proses penelitian yang telah dilakukan yang mencakup implikasi dari hasil pemodelan topik dalam bentuk skripsi.

2.2 Waktu dan Lokasi Penelitian

Penelitian dilakukan sejak disetujuinya proposal penelitian pada tanggal 16 Mei 2024. Penelitian dilakukan di Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin.

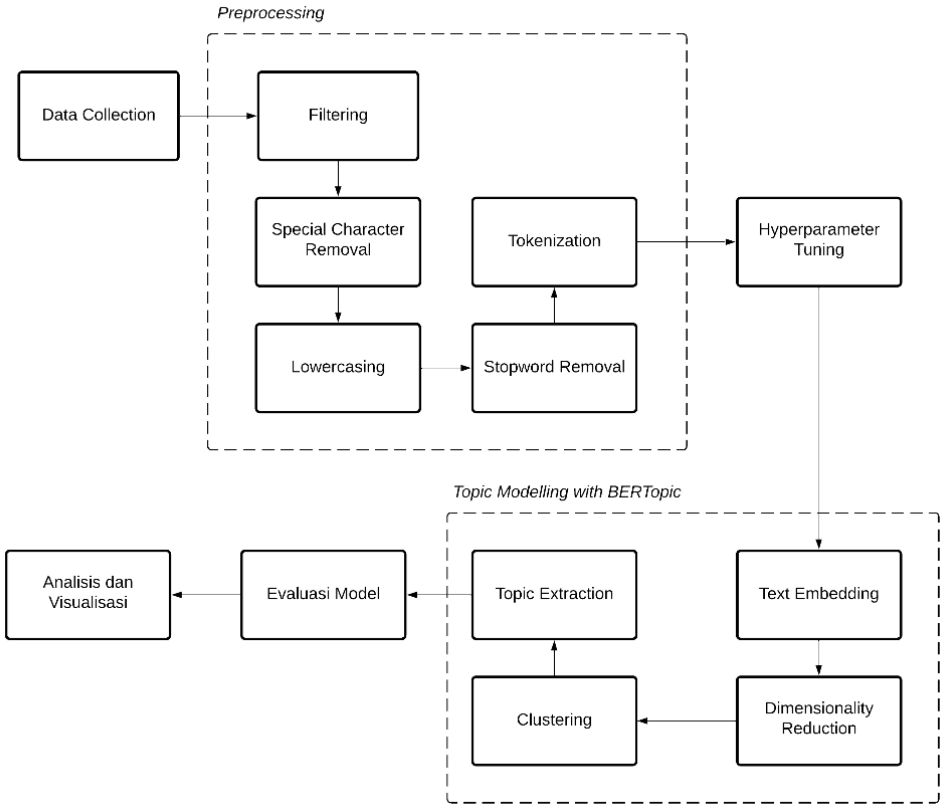
2.3 Instrumen Penelitian

Alat yang digunakan dalam melaksanakan penelitian ini terdiri dari perangkat keras dan perangkat lunak sebagai berikut:

1. Perangkat keras
   a. Laptop *Lenovo Legion 5* dengan *Processor AMD Ryzen 5*, RAM 16GB, SSD 512 GB
2. Perangkat lunak
   a. Windows 11 x64
   b. Visual Studio Code
   c. *Python 3.10.6*

2.4 Rancangan Sistem

Bagian ini menjelaskan rancangan sistem yang digunakan dalam penelitian ini. Rancangan sistem disusun berdasarkan proses pengerjaan yang digambarkan dalam bentuk diagram alur.



Gambar 2. Diagram Alur Rancangan Sistem Penelitian

2.4.1 Data Collection

Abstrak merupakan bagian dari publikasi ilmiah yang digunakan sebagai data dalam penelitian ini. Abstrak adalah rangkuman dari isi dokumen atau publikasi ilmiah yang menyajikan informasi inti, seperti tujuan penelitian, metode, hasil, dan kesimpulan. Dalam penelitian ini, abstrak diambil dari repositori arXiv. ArXiv dipilih sebagai sumber dataset karena berbagai hal yang mendukung kualitas dan efisiensi proses pengumpulan data. Struktur situs yang sederhana dan minim *noise* akan mempermudah dan meminimalkan kesalahan dalam pengumpulan data. Selain itu, arXiv merupakan repositori akses terbuka yang menyediakan publikasi berkualitas dari berbagai disiplin ilmu yang berasal dari peneliti di berbagai institusi terkemuka di dunia dengan pembaruan yang berkala.

Data akan diambil menggunakan metode *web scraping* secara automatis dengan mengumpulkan informasi yang diperlukan seperti *id* publikasi ilmiah, tanggal publikasi dan terutama abstrak dari setiap publikasi yang diterbitkan dalam kategori *Computer*

Science. Data yang akan diproses merupakan abstrak yang diterbitkan per Januari 2023 sampai Desember 2023. Proses ini dilakukan dengan bantuan *library* python *BeautifulSoup4*.

2.4.2 Preprocessing

Preprocessing data merupakan tahapan untuk membersihkan dan mempersiapkan teks sebelum digunakan dalam pemodelan topik. Teknik *prepocessing* yang dilakukan dalam penelitian ini adalah sebagai berikut.

2.4.2.1 Filtering

Data yang digunakan pada penelitian ini difilter sehingga data hanya mencakup publikasi yang diterbitkan pada tahun 2023. Filtering dilakukan dengan mengolah data yang mengandung informasi tanggal dikonversi menjadi format yang relevan kemudian menghapus data diluar dari tahun 2023 sesuai dengan tanggal *submitted* pada arXiv. Hal ini dilakukan untuk memastikan kesesuaian data dengan periode yang dianalisis.

2.4.2.2 Special Character Removal

Menghapus karakter khusus (misalnya, tanda baca) yang tidak memberikan kontribusi signifikan dalam analsis teks. Karakter seperti tanda baca dan karakter spesial lain (%, &, *, @, #, dan lain-lain) akan dihilangkan untuk membersihkan teks dan hanya mempertahankan data alfanumerik dan spasi antar kata.

2.4.2.3 Lowercasing

Mengubah semua huruf dalam teks menjadi huruf kecil untuk memastikan konsistensi dalam analisis terutama duplikasi kata yang sama dalam bentuk huruf besar dan huruf kecil (misalnya, kata "*Computer*" dan "*computer*").

2.4.2.4 Stopword Removal

Menghapus kata-kata umum atau *stopword* (dalam bahasa Inggris) yang tidak memiliki arti penting (misalnya, "*I*", "*you*", "*she*", "*are*", "*who*", dan "*been*") untuk meningkatkan kualitas analisis teks. Proses ini bertujuan untuk meningkatkan kualitas data dan efisiensi model salama proses pembelajaran mesin.

2.4.2.5 Tokenisasi

Tokenisasi adalah proses membagi teks atau dokumen menjadi bagian-bagian kecil seperti kalimat utuh, frasa, atau kata yang disebut *token*. Tokenisasi yang digunakan dalam penelitian ini adalah tokenisasi ke bentuk kata. Contoh kalimat "*A Compact Introduction to Fractional Calculus*" dipecah menjadi ("*A", "Compact", "Introduction", "to", "Fractional", "Calculus*").

2.4.3 Hyperparameter Tuning

Hyperparameter tuning dilakukan untuk mendapatkan hasil yang diharapkan dalam pemodelan topik. Terdapat parameter tetap dan parameter yang disesuaikan untuk masing-masing dataset. Metode eksperimental *grid search* digunakan untuk mengeksplorasi kombinasi optimal dari parameter-parameter tersebut, dengan mempertimbangkan keseimbangan antara *topic coherence* dan jumlah topik yang dihasilkan.

2.4.4 Pemodelan dengan BERTopic

Proses pemodelan topik dalam penelitian ini dilakukan menggunakan *BERTopic*. *BERTopic* memliki sistem yang bekerja melalui empat tahapan utama, mulai dari *text embedding*, pengurangan dimensi, *clustering* dan ekstraksi topik.

2.4.4.1 Text Embedding

Proses *text embedding* dilakukan untuk merepresentasikan suatu kata atau kalimat ke dalam bentuk *dense vector*. Proses ini menggunakan *pre-trained sentence transformer*. Adapun *input* data pada proses ini yaitu teks abstrak yang telah dikumpulkan. Hal ini dilakukan dengan tujuan mengubah data teks menjadi representasi numerik.

2.4.4.2 Pengurangan Dimensi dengan UMAP

Proses *text embedding* akan meningkatkan dimensi data, sehingga perlu dilakukan pengurangan dimensi menggunakan UMAP (*Uniform Manifold Approximation and Projection*). UMAP menunjukkan lebih banyak fitur lokal dan global data pada data berdimensi tinggi yang diproyeksikan ke dimensi yang lebih rendah (McInnes et al., 2020). Input yang digunakan pada proses ini berupa representasi numerik yang dihasilkan dari proses *text embedding*.

2.4.4.3 Clustering dengan HDBSCAN

Pada pemodelan topik menggunakan *BERTopic,* proses *clustering* dari hasil *text embedding* yang telah melalui proses pengurangan dimensi menggunakan UMAP, dilakukan menggunakan metode HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). Metode HDBSCAN menggunakan pendekatan *soft clustering*, dimana *noise* dimodelkan sebagai *outlier*. Input dari proses ini merupakan representasi numerik dari hasil pengurangan dimensi oleh UMAP.

2.4.4.4 Ekstraksi Topik dengan c-TF-IDF

Setelah melalui proses *clustering,* c-TF-IDF (*Class-Based Term Frequency-Inverse Document Frequency*) digunakan untuk menghasilkan representasi topik dari tiap *cluster* yang dihasilkan. Penggunaan c-TF-IDF menghasilkan kata-kata teratas atau *top words* yang merepresentasikan setiap topik atau *cluster* dengan memodelkan pentingnya suatu kata dalam *cluster*.

2.4.4.5 Metode Interpretasi Topik

Proses interpretasi topik dilakukan dengan menggunakan *top words* (kata-kata teratas) yang dihasilkan model sebagai acuan. Untuk memastikan relevansi dan validitas interpretasi, kami memanfaatkan taksonomi dari organisasi profesional internasional di bidang teknologi dan ilmu komputer, yaitu IEEE (*Institute of Electrical and Electronics Engineers*, 2025) . Dengan demikian, hasil interpretasi tidak hanya bergantung pada asumsi yang bersifat subjektif, tetapi juga didasari referensi yang diakui secara internasional.

2.4.5 Evaluasi Model

Pada tahap ini, model yang telah dihasilkan dievaluasi untuk mengukur kualitas topik yang dihasilkan oleh *BERTopic* menggunakan metrik *c_v*. Metrik ini digunakan untuk mengukur seberapa baik interpretabilitas dan keterkaitan semantik topik yang dihasilkan

model. Nilai *coherence score* berkisar antara 0 hingga 1 di mana nilai yang mendekati 1 menunjukkan topik yang lebih koheren dan bermakna. Perhitungan *coherence* dilakukan berdasarkan pendekatan yang diusulkan Röder et al. (2015). Tahapan ini diterapkan menggunakan *library Gensim*, yang menyediakan fungsi *CoherenceModel* untuk membantu menghitung *coherence score*.

2.4.6 Analisa Hasil

ArXiv telah membagi ilmu komputer atau *Computer Science* menjadi 40 kategori. Analisis dilakukan dengan memetakan topik-topik yang dihasilkan model ke dalam kategori-kategori yang relevan, kemudian membandingkan proporsi dan pola trennya dengan data aktual dari arXiv. Analisis ini dilakukan untuk melihat seberapa baik *BERTopic* dapat merepresentasikan distribusi dokumen dari setiap kategori sesuai dengan data aktualnya. Proporsi setiap kategori dihitung dengan menggunakan rumus:

$$Proporsi = \frac{Jumlah\ dokumen\ dalam\ kategori}{Total\ jumlah\ dokumen} \times 100\% \qquad (1)$$

Dalam rumus ini, perhitungan proporsi data aktual arXiv dan hasil *BERTopic* memiliki perbedan dalam menentukan jumlah dokumen. Untuk data aktual arXiv, jumlah dokumen dalam kategori merujuk pada jumlah dokumen dari kategori tertentu di arXiv, sedangkan total jumlah dokumen adalah akumulasi jumlah dokumen dari setiap kategori. Untuk perhitungan proporsi hasil *BERTopic*, jumlah dokumen dalam kategori merujuk pada jumlah dokumen dari masing-masing topik yang dipetakan ke dalam kategori yang relevan dengan topik-topik tersebut.

Untuk membandingkan pola tren, jumlah dokumen arXiv per bulan dari setiap kategori dikumpulkan kemudian dibandingkan dengan jumlah dokumen hasil pemetaan topik per bulan untuk setiap kategori.

Melalui proses ini, diharapkan dapat memperoleh pemahaman mendalam mengenai kemampuan *BERTopic* dalam menangani data teks dari abstrak publikasi ilmiah dan membantu mengetahui apakah terdapat kesalahan dalam pengelompokan dokumen atau topik-topik yang tumpang tindih.