

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era digital saat ini, kemajuan teknologi informasi mempermudah akses ke berbagai sumber publikasi ilmiah yang sangat dibutuhkan oleh civitas akademika, baik mahasiswa maupun dosen, untuk mendukung penelitian, penulisan skripsi, maupun pengembangan akademik. Teknologi yang semakin canggih ini membuat akses ke jurnal, artikel, dan portal data makin luas dan mudah sehingga informasi dan referensi bisa diperoleh dengan instan dalam aktivitas kampus sehari-hari (Anggraini, 2020). Publikasi ilmiah kini menjadi acuan penting, baik dalam penyusunan laporan penelitian, proposal riset, maupun tugas akhir mahasiswa. Meski demikian, kemudahan akses ini justru memunculkan tantangan baru, yaitu banyaknya pilihan referensi yang membuat pengguna terkadang kesulitan dalam menentukan publikasi yang sesuai kebutuhan atau minat penelitian. Akibatnya, proses pencarian dan pemilahan topik sering memakan waktu lama serta menyulitkan pemetaan kecenderungan tema di lingkungan akademik (Titiana & Bangkalang, 2023). Situasi ini juga dialami oleh mahasiswa dan dosen di lingkungan Pendidikan Teknik Informatika dan Komputer (PTIK) Universitas Negeri Jakarta (UNJ), yang kian tahun produktivitas artikel dan publikasinya semakin bertambah.

Program Studi Pendidikan Teknik Informatika dan Komputer Universitas Negeri Jakarta menaungi beragam penelitian dalam bidang Rekayasa Perangkat Lunak (RPL), Multimedia (MM), dan Teknik Komputer Jaringan (TKJ), yang masing-masing memiliki karakter dan isu keilmuan tersendiri. Jurnal Pendidikan Teknik Informatika dan Komputer (PINTER) sebagai wahana utama publikasi penelitian di lingkungan PTIK UNJ, memuat luaran riset dari ketiga peminatan tersebut secara periodik dan berkesinambungan. Setiap edisi jurnal menghimpun karya ilmiah dari berbagai spektrum tema, mulai dari pengembangan aplikasi hingga multimedia edukasi dan sistem jaringan digital. Meski demikian, belum ada mekanisme identifikasi kecenderungan tema penelitian secara otomatis dan terukur yang diadopsi pengelola jurnal maupun civitas akademika. Sebagian besar proses

rekapitulasi dan analisis topik penelitian masih dilakukan secara manual, yang rentan subyektivitas dan kelemahan dalam aspek kecepatan serta keakuratan pemetaan.

Seiring dengan meningkatnya volume data artikel ilmiah yang diterbitkan setiap tahun, kebutuhan akan metode pemetaan topik secara otomatis dan sistematis menjadi semakin mendesak. Kondisi ini juga tercermin pada perkembangan Jurnal PINTER yang menunjukkan tren peningkatan jumlah publikasi setiap tahunnya. Pada tahun 2024, misalnya, Jurnal PINTER menerbitkan dua volume, yaitu Volume 8 Nomor 1 dan Volume 8 Nomor 2, dengan masing-masing volume berisi 10 artikel, sehingga total artikel yang terbit dalam satu tahun mencapai 20 (PINTER, 2025).

Peningkatan jumlah publikasi tersebut menimbulkan tantangan baru dalam mengelola dan menganalisis konten artikel, khususnya dalam hal identifikasi tema penelitian yang relevan. Metode konvensional seperti membaca judul dan abstrak secara manual menjadi kurang efisien, terutama karena banyak artikel memiliki subjek yang tumpang tindih atau bersifat multidisipliner. Tanpa adanya sistem analisis otomatis, mahasiswa kerap mengalami kesulitan dalam memilih referensi dan menentukan topik penelitian yang sesuai dengan kebutuhan dan tren aktual. Kondisi ini juga berdampak pada dosen pembimbing serta pengelola jurnal dalam menelusuri kontribusi penelitian dan menentukan arah pengembangan riset. Ketidakefektifan dalam pengelompokan topik dan identifikasi struktur keilmuan dapat berpengaruh terhadap penyusunan peta jalan riset (*roadmap*) yang efektif di tingkat program studi maupun fakultas (Rizki Ramadandi, 2021).

Pendekatan otomatis seperti *topic modeling* menjadi solusi utama dalam mengelompokkan ribuan dokumen secara efisien berdasarkan kemiripan tema teks yang tersembunyi. *Topic modeling* yang berbasis pada teknik *text mining* mampu menemukan pola keterhubungan kata dan tema secara matematis, tanpa memerlukan anotasi manual. Salah satu algoritma *topic modeling* yang paling banyak diadopsi dan dibuktikan ampuh adalah *Latent Dirichlet Allocation* (LDA) yang dikembangkan Blei dan kolega pada 2003. LDA sendiri menggunakan model probabilistik generatif sehingga setiap dokumen dapat mengandung lebih dari satu tema dengan proporsi tertentu berdasarkan distribusi kata. Dengan kelebihan ini, LDA membawa perubahan signifikan dalam pemetaan tema penelitian ilmiah,

terutama untuk kebutuhan klasifikasi otomatis publikasi di jurnal-jurnal akademik modern (Blei et al., 2003).

Penerapan LDA berhasil dibuktikan dalam sejumlah penelitian terbaru, baik pada jurnal kesehatan, Pendidikan, maupun bidang komputer. Studi Titiana & Bangkalang (2023) memperlihatkan hasil identifikasi otomatis empat cluster topik utama tugas akhir mahasiswa dengan *coherence score* mencapai 0,61, mengindikasikan akurasi topik yang interpretatif secara semantik. Penelitian lain oleh Herwanto (2018) menunjukkan integrasi LDA dengan machine learning berbasis clustering mampu mereduksi dimensi representasi dokumen, sehingga distribusi tema lebih mudah dijelaskan dan divisualisasikan (Herwanto, 2018). Hal serupa juga dicapai oleh Ramadandi (2021), yang membandingkan akurasi LDA dan *Term Frequency – Inverse Document Frequency* (TF-IDF), serta membuktikan keunggulan LDA dalam menangkap variasi topik pada kumpulan dokumen ilmiah skala besar (Rizki Ramadandi, 2021). Kualitas pemodelan biasanya dievaluasi menggunakan *coherence score* sehingga hasilnya tidak hanya signifikan secara matematis, tetapi juga bermakna secara semantik sesuai kebutuhan pemetaan ilmu terkini (Titiana & Bangkalang, 2023).

Meskipun LDA telah banyak diterapkan, kinerjanya sangat dipengaruhi oleh representasi fitur yang digunakan sebagai masukan model. TF-IDF merupakan representasi fitur yang umum digunakan karena mampu menonjolkan kata-kata yang secara statistik penting dan khas pada suatu dokumen berdasarkan frekuensi kemunculannya, namun pendekatan ini tidak mempertimbangkan hubungan semantik antar kata sehingga kata-kata yang berbeda bentuk namun memiliki makna serupa akan diperlakukan sebagai unit yang sepenuhnya berbeda. Di sisi lain, *FastText* sebagai metode *word embedding* mampu menangkap keterkaitan makna antar kata melalui representasi vektor, namun tidak mempertimbangkan seberapa penting suatu kata dalam membedakan satu dokumen dari dokumen lain. Kombinasi antara TF-IDF dan *FastText* berpotensi mengatasi keterbatasan masing-masing representasi tersebut, sehingga dapat menghasilkan topik yang lebih koheren dan representatif. Namun, penelitian-penelitian terdahulu yang disebutkan di atas umumnya masih menggunakan representasi fitur secara tunggal, sehingga belum diketahui sejauh mana kombinasi TF-IDF dan *FastText* dapat meningkatkan

kualitas hasil pemodelan topik menggunakan LDA, khususnya pada korpus publikasi ilmiah berbahasa Indonesia seperti Jurnal PINTER. Selain itu, evaluasi kualitas model topik tidak hanya dapat dilihat dari sisi keterpahaman makna melalui *coherence score*, tetapi juga perlu didukung oleh metrik *perplexity* yang menggambarkan kemampuan model dalam memprediksi distribusi kata pada data, sehingga kedua metrik tersebut perlu digunakan secara bersamaan untuk memperoleh gambaran kualitas model yang lebih utuh.

Berdasarkan berbagai temuan dan metodologi penelitian lima tahun terakhir, pendekatan LDA terbukti unggul tidak hanya untuk pemetaan topik, namun juga mendukung pengembangan *roadmap* penelitian dan pengelolaan data akademik (Rizki Ramadandi, 2021; Titiana & Bangkalang, 2023). Penelitian ini akan mengaplikasikan LDA pada seluruh artikel di jurnal PINTER sebagai basis pemetaan tema riset masa kini sehingga akan memudahkan mahasiswa maupun dosen dalam menemukan kecenderungan tema sekaligus peluang riset berikutnya, sesuai dengan kebutuhan akademik dan tantangan teknologi terbaru. Penerapan metode ini sekaligus dapat menjadi referensi pengelola jurnal dalam penyusunan strategi publikasi serta menjaga kualitas dan arah topik penelitian di bidang Pendidikan teknik informatika dan komputer.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, terdapat beberapa permasalahan yang dapat diidentifikasi dalam penelitian ini, yaitu:

1. Banyaknya artikel dalam jurnal PINTER yang memiliki beragam tema kajian menyebabkan kesulitan dalam mengidentifikasi topik utama secara manual.
2. Belum terdapat mekanisme otomatis untuk pemetaan topik di jurnal PINTER secara komputasi berbasis model statistik.
3. Belum ada penelitian yang menganalisis tema jurnal PINTER secara mendalam menggunakan pendekatan *text mining* berbasis *topic modeling*.
4. Penelitian sebelumnya umumnya hanya menggunakan satu jenis representasi fitur, sehingga belum diketahui pengaruh kombinasi TF-IDF dan *FastText* terhadap kualitas topik yang dihasilkan oleh metode *Latent Dirichlet Allocation* (LDA).

1.3 Pembatasan Masalah

Agar penelitian ini lebih terarah dan tidak menyimpang dari tujuan yang telah ditetapkan, maka diperlukan beberapa batasan masalah, yaitu:

1. Penelitian dilakukan pada artikel yang dipublikasi di jurnal PINTER dari volume 1 seri 1 hingga volume 8 seri 2.
2. Pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA).
3. Analisis dilakukan hanya pada teks utama artikel, tanpa mempertimbangkan metadata, tabel, gambar, dan elemen non-teks lainnya.

1.4 Perumusan Masalah

Berdasarkan latar belakang, identifikasi masalah, dan batasan masalah yang telah diuraikan, maka perumusan masalah dalam penelitian ini adalah: Bagaimana hasil evaluasi *topic modeling* pada jurnal PINTER menggunakan metode *Latent Dirichlet Allocation* (LDA)?

1.5 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengevaluasi kualitas model *topic modeling* pada Jurnal PINTER menggunakan metode *Latent Dirichlet Allocation* (LDA) dengan representasi TF-IDF, *FastText*, dan kombinasi TF-IDF + *FastText* berdasarkan metrik *Coherence Score* dan *Perplexity*, untuk menentukan representasi fitur yang menghasilkan kualitas *topic modeling* terbaik.

1.6 Manfaat Penelitian

Penelitian ini memiliki dua jenis manfaat utama, yaitu manfaat teoritis dan manfaat praktis. Kedua manfaat ini diharapkan dapat memberikan kontribusi baik bagi pengembangan ilmu pengetahuan maupun bagi pihak-pihak yang berkepentingan secara langsung.

1.6.1 Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan manfaat secara teoritis, di antaranya:

1. Memberikan kontribusi dalam pengembangan kajian text mining khususnya penerapan metode Latent Dirichlet Allocation (LDA) pada publikasi ilmiah berbahasa Indonesia.
2. Menambah referensi empiris mengenai efektivitas *coherence score* sebagai indikator evaluasi kualitas pemodelan topik.
3. Memperkaya literatur di bidang Pendidikan Teknik Informatika dan Komputer terkait pemetaan topik penelitian berbasis analisis kuantitatif.

1.6.2 Manfaat Praktis

Selain manfaat teoritis, penelitian ini juga diharapkan dapat memberikan manfaat praktis bagi beberapa pihak, yaitu:

Bagi Pengelola Jurnal PINTER:

1. Menyediakan informasi objektif mengenai distribusi dan arah topik penelitian yang dipublikasikan dalam jurnal.
2. Menjadi dasar dalam merancang strategi penerbitan serta pengembangan fokus tema penelitian di masa mendatang.

Bagi Mahasiswa:

1. Menyediakan referensi empiris mengenai perkembangan riset di bidang Pendidikan Teknik Informatika dan Komputer.
2. Membantu mahasiswa dalam menentukan topik penelitian yang relevan, terkini, dan sesuai dengan tren penelitian di lingkungan program studi.
3. Memberikan contoh penerapan algoritma Latent Dirichlet Allocation dan evaluasi *coherence score* dan *perplexity* dalam konteks analisis publikasi ilmiah berbahasa Indonesia, sehingga dapat dijadikan acuan untuk penelitian lanjutan di bidang text mining dan machine learning.