BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi telah mengubah cara informasi ilmiah diakses dan dikelola, terutama dalam publikasi akademik. Saat ini, akses ke sumber ilmiah semakin mudah melalui *platform* seperti *Google Scholar, Scopus*, dan repositori universitas. Bagi dosen dan mahasiswa, publikasi ilmiah menjadi dasar penting untuk menyusun skripsi, tesis, disertasi, dan proposal penelitian. Namun, meskipun data digital tersedia melimpah, menemukan topik yang relevan masih menjadi tantangan. Banyak pengguna kesulitan menentukan kategori atau arah penelitian, terutama jika mereka kurang paham tentang bidang tertentu. Hal ini dapat memperlambat proses penelitian, menyebabkan duplikasi, atau mengarahkan pada pemilihan topik yang kurang tepat.

Di Universitas Negeri Jakarta (UNJ), masalah ini juga terjadi. Sebagai universitas yang fokus pada pendidikan dan penelitian, UNJ memiliki banyak dosen yang aktif menerbitkan karya ilmiah. Bidang komputer, khususnya di program studi (prodi) Ilmu Komputer, Pendidikan Teknik Informatika dan Komputer, serta Sistem dan Teknologi Informasi, berkembang pesat. Dosen-dosen di prodi ini telah menghasilkan banyak publikasi yang terindeks di *Scopus*. Namun, belum ada sistem otomatis yang dapat mengelompokkan publikasi tersebut berdasarkan topik bidang komputer secara spesifik. Akibatnya, pemetaan keahlian dosen menjadi sulit dilakukan, dan mahasiswa atau peneliti lain kesulitan menemukan ahli di sub-bidang tertentu.

Secara internasional, *Association for Computing Machinery* (ACM) telah mengembangkan ACM *Computing Classification System* (ACM CCS) sebagai standar taksonomi dalam ilmu komputer untuk mengorganisir lanskap disiplin komputasi yang luas, yang secara resmi

mencakup bidang-bidang seperti *Computer Engineering* (CE), *Computer Science* (CS), *Cybersecurity* (CSEC), *Information Systems* (IS), *Information Technology* (IT), *Software Engineering* (SE), dan *Data Science* (DS). Dengan semakin banyaknya jumlah topik penelitian, penataan secara hierarkis menjadi sebuah kebutuhan (Sadat & Caragea, 2022). ACM CCS menyediakan struktur hierarki bertingkat seringkali hingga enam level yang memungkinkan sebuah karya ilmiah diklasifikasikan tidak hanya pada satu topik spesifik, tetapi juga pada seluruh kategori induknya hingga ke akar (Sadat & Caragea, 2022).

Upaya untuk mengekstrak keahlian dosen secara otomatis dari publikasi ilmiah telah terbukti dapat dilakukan dengan tingkat keberhasilan yang meyakinkan. Salah satu bukti nyata ditunjukkan oleh penelitian Aini & Yulianti (2025) yang sukses mengembangkan sistem pemrofilan kepakaran dosen dengan mengacu pada standar taksonomi ACM CCS. Melalui eksperimen yang menggabungkan fitur *adjusted* TF-IDF dengan berbagai metode *embedding* (seperti SBERT, mBERT, XLM-R, dan FastText), sistem tersebut mampu memetakan judul dan abstrak publikasi secara akurat dengan nilai P@1 mencapai 0,75. Pencapaian ini menjadi bukti empiris yang kuat bahwa standar ACM sangat efektif digunakan sebagai acuan identifikasi kepakaran secara otomatis. Kesuksesan kerangka pemetaan inilah yang menjadi landasan utama untuk diimplementasikan pada publikasi dosen bidang komputer di Universitas Negeri Jakarta (UNJ).

Penelitian sebelumnya dalam klasifikasi teks seringkali mengandalkan metode statistik seperti *K-Means Clustering* atau pemodelan topik seperti *Latent Dirichlet Allocation* (LDA) (Septianda et al., 2025). Meskipun metode-metode ini dapat mengidentifikasi kelompok kata, mereka memiliki keterbatasan fundamental, terutama dalam menangani teks pendek seperti judul publikasi. Metode ini cenderung mengabaikan urutan kata dan konteks sekuensial, sehingga

kurang mampu menangkap makna semantik yang mendalam yang krusial untuk klasifikasi akurat.

Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan pendekatan *deep learning* yang mampu memahami data sekuensial. Model *Long Short-Term Memory* (LSTM) dan variannya, *Bidirectional Long Short-Term Memory* (BiLSTM), merupakan arsitektur *Recurrent Neural Network* (RNN) yang terbukti unggul dalam pemrosesan bahasa alami. LSTM dirancang khusus untuk mengatasi masalah *gradient vanishing* yang sering terjadi pada RNN standar (Jasmir et al., 2022).Sementara itu, BiLSTM menyempurnakan LSTM dengan memproses data dari dua arah (maju dan mundur), memungkinkannya untuk menangkap informasi kontekstual dari masa lalu dan masa depan secara simultan (He et al., 2024). Kemampuan ini menjadikannya model yang sangat efektif untuk tugas klasifikasi teks, karena dapat memanfaatkan susunan kata secara lebih komprehensif untuk mengekstrak fitur semantik

Untuk mengubah teks menjadi representasi numerik yang padat makna sebelum diproses oleh model LSTM dan BiLSTM, penelitian ini akan memanfaatkan teknik *word embedding*. Ini adalah teknik representasi kata dalam bentuk vektor berdimensi rendah di mana kata-kata dengan makna serupa akan memiliki vektor yang berdekatan (Johnson et al., 2023; Camacho-Collados & Pilehvar, 2021). Penelitian ini akan mengeksplorasi dua metode utama: Word2Vec dan FastText. Word2Vec dikenal efektif dalam menangkap hubungan semantik berbasis konteks (Camacho-Collados & Pilehvar, 2018; Apidianaki, 2022), sementara FastText dilibatkan untuk mengatasi keterbatasan umum Word2Vec, terutama kemampuannya menangani kata-kata di luar kamus (*Out-of-Vocabulary*/OOV) dan informasi morfologi melalui mekanisme *sub-word* (Choi & Lee, 2020; Tulu, 2022). Pendekatan ini dipilih karena *word embedding* terbukti secara substansial meningkatkan kinerja model klasifikasi, seperti yang ditunjukkan oleh Jasmir et al., 2022 pada model BiLSTM.

Meskipun model *deep learning* LSTM dan BiLSTM akan menjadi metode klasifikasi utama yang bersifat terawasi (*supervised*), penelitian ini juga akan memanfaatkan *Latent Dirichlet Allocation* (LDA) sebagai metode pendukung yang tidak terawasi (*unsupervised*). Peran LDA di sini bukanlah sebagai alat klasifikasi akhir, melainkan sebagai alat bantu untuk analisis topik awal. LDA bekerja dengan mengidentifikasi gugusan kata yang sering muncul bersama di dalam korpus publikasi untuk membentuk "topik" laten tanpa memerlukan label kategori yang telah ditentukan sebelumnya (Asudani et al., 2023). Hasil dari LDA ini akan memberikan wawasan awal mengenai tema-tema penelitian dominan yang ada di UNJ. Wawasan ini kemudian dapat digunakan untuk membantu dan memvalidasi proses pelabelan data sesuai dengan standar ACM CCS, sehingga memastikan konsistensi dan relevansi label yang akan digunakan untuk melatih model LSTM dan BiLSTM

Secara keseluruhan, penelitian ini bertujuan mengembangkan sistem klasifikasi otomatis untuk publikasi ilmiah bidang komputer di UNJ. Dengan mengadopsi ACM CCS sebagai standar, penelitian ini mengusulkan model klasifikasi berbasis *deep learning* (LSTM dan BiLSTM) yang didukung oleh teknik *word embedding* (Word2Vec dan FastText) untuk representasi fitur yang kuat. Sebagai analisis pelengkap, metode LDA digunakan untuk melakukan pemodelan topik guna mendukung proses pelabelan data. Kombinasi pendekatan ini diharapkan dapat menghasilkan sistem yang akurat dalam memetakan kepakaran dosen dan memberikan wawasan mendalam mengenai lanskap penelitian di UNJ.

1.2 Identifikasi Masalah

Berdasarkan latar belakang, masalah yang ditemukan adalah:

1. Belum adanya sistem otomatis untuk mengelompokkan publikasi dosen bidang komputer di UNJ (Prodi Ilmu Komputer, PTIK, dan STI) menurut standar ACM CCS berdasarkan data dari Scopus.

2. Adanya kesulitan dalam menangkap nuansa makna dan konteks pada judul. Abstrak, kata kunci dan metodologi publikasi ilmiah, yang menyebabkan klasifikasi menjadi tidak akurat jika menggunakan metode konvensional.
3. Peta keahlian dan tren riset para dosen bidang komputer di UNJ belum teridentifikasi secara jelas dan terstruktur sehingga menyulitkan proses penemuan pakar.
4. Ketergantungan data latih berlabel untuk model *deep learning* (LSTM dan BiLSTM), sementara proses pelabelan manual ribuan abstrak, kata kunci dan metodologi ke dalam kategori ACM CCS merupakan tantangan yang kompleks dan memakan waktu.

1.3 Pembatasan Masalah

Agar penelitian lebih terfokus, penelitian ini dibatasi pada:

1. Data yang digunakan meliputi atribut judul, abstrak, metodologi, dan kata kunci dari publikasi ilmiah dosen Program Studi Ilmu Komputer, Pendidikan Teknik Informatika dan Komputer, serta Sistem dan Teknologi Informasi UNJ yang terindeks Scopus, dengan rentang waktu publikasi mulai dari data tersedia (*all years*) hingga Desember 2025.
2. Pendekatan klasifikasi utama menggunakan model *deep learning Long Short-Term Memory* (LSTM) dan *Bidirectional Long Short-Term Memory* (BiLSTM), serta didukung oleh *word embedding* Word2Vec dan FastText.
3. Kategori klasifikasi mengacu pada standar ACM Computing Classification System (ACM CCS).
4. Metode *Latent Dirichlet Allocation* (LDA) digunakan sebagai analisis komplementer untuk mengidentifikasi klaster topik secara *unsupervised* yang hasilnya dapat menjadi pertimbangan dalam proses pelabelan data latih.
5. Penelitian ini hanya berfokus pada tahap eksperimen, pembangunan, dan evaluasi performa model klasifikasi, serta tidak

mencakup implementasi model ke dalam bentuk aplikasi perangkat lunak, sistem informasi, atau antarmuka pengguna (user interface).

1.4 Rumusan Masalah

Berdasarkan masalah yang ada, pertanyaan utama penelitian ini adalah: "Bagaimana perbandingan performa model *deep learning* LSTM dan BiLSTM dalam mengklasifikasikan publikasi ilmiah dosen bidang komputer UNJ berdasarkan standar ACM CCS?"

1.5 Tujuan Penelitian

Tujuan penelitian ini adalah menganalisis dan membandingkan performa model klasifikasi menggunakan LSTM dan BiLSTM untuk mengelompokkan publikasi dosen bidang komputer UNJ sesuai standar ACM CCS, serta menentukan model yang paling efektif yang didukung oleh analisis topik dari LDA untuk membantu proses pengelompokan.

1.6 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah:

1. Bagi institusi akademik: Memberikan data klasifikasi topik yang akurat untuk mendukung perencanaan strategis penelitian, evaluasi kinerja, dan proses akreditasi.
2. Bagi Dosen dan mahasiswa: Menyediakan sistem untuk memetakan kepakaran dosen dan tren riset, mempermudah pencarian kolaborator atau pembimbing yang relevan.
3. Bagi pengembangan ilmu: Menawarkan evaluasi komparatif antara model LSTM dan BiLSTM dalam konteks klasifikasi publikasi ilmiah berbahasa Indonesia, yang dapat menjadi referensi bagi penelitian sejenis