ABSTRAK

Dinamika industri teknologi informasi (IT) di Indonesia menuntut adaptasi kualifikasi keahlian teknis yang masif. Ketidakmampuan pencari kerja dan institusi pendidikan dalam memenuhi kebutuhan riil industri ini pada akhirnya menciptakan *skill gap* yang signifikan. Upaya pemetaan kebutuhan industri sebelumnya sering menggunakan pendekatan *supervised learning*, namun metode ini cenderung statis dan kaku dalam menghadapi pola kemunculan keahlian teknis baru akibat ketergantungannya pada data berlabel. Menanggapi keterbatasan tersebut, penelitian ini mengangkat rumusan masalah tentang bagaimana membangun sistem yang mampu mengidentifikasi dan memetakan pola kebutuhan keahlian teknis secara otomatis. Melalui ekstraksi dokumen lowongan pekerjaan, penelitian ini bertujuan untuk menghasilkan model yang dapat memetakan pilar spesialisasi keahlian yang secara nyata mendominasiw pasar kerja.

Penelitian ini menerapkan metode kuantitatif deskriptif *Natural Language Processing* (NLP) dengan kerangka kerja CRISP-DM. Proses pengumpulan data dilakukan melalui teknik *web scraping* pada portal LinkedIn wilayah Indonesia selama periode kuartal keempat tahun 2025. Dengan mengedepankan etika pengambilan data publik, ekstraksi dijalankan secara otomatis menggunakan skema tanpa *login* yang didukung oleh pustaka Selenium dan BeautifulSoup, serta berhasil mengumpulkan 8.507 baris data mentah. Selanjutnya, data tersebut melewati rangkaian prapemrosesan (*preprocessing*) yang komprehensif—meliputi *deduplication*, pembersihan data, *case folding*, normalisasi, penanganan kata majemuk (*compound terms handling*), tokenisasi, penyaringan dokumen, hingga *whitelist-based filtering* untuk mendapatkan *hard skill* secara murni. Setelah teks bersih diubah menjadi matriks *Bag-of-Words* (BoW), algoritma *unsupervised learning* berupa *Latent Dirichlet Allocation* (LDA) diaplikasikan untuk memetakan klaster keahlian dengan cara mengungkap struktur semantik tersembunyi dari probabilitas kemunculan antar kata.

Algoritma *Latent Dirichlet Allocation* (LDA) terbukti efektif dalam memetakan spesialisasi *skill* IT di Indonesia. Evaluasi model menunjukkan performa terbaik pada konfigurasi lima topik (K=5), dengan *Coherence Score* 0.5637 (menunjukkan kepaduan semantik yang kuat) dan *Topic Diversity* 0.736 (pemisahan klaster yang jelas tanpa tumpang tindih). Penelitian ini sukses menjawab permasalahan utama dengan mengekstraksi lima pilar keahlian dominan: *Frontend & UI/UX Development*, *Cloud & DevOps Engineering*, *Data Science & Machine Learning*, *Backend Development*, dan *Mobile Development*, di mana *Frontend* menempati peringkat teratas dengan proporsi permintaan sebesar 36,62%. Luaran penelitian ini berkontribusi sebagai landasan empiris bagi institusi pendidikan tinggi untuk mengevaluasi kurikulum pembelajaran, sekaligus menjadi kompas strategis bagi pencari kerja dalam menyelaraskan keterampilan mereka dengan kualifikasi riil yang diserap industri teknologi.

Kata Kunci: Lowongan Pekerjaan, Pemodelan Topik, *Latent Dirichlet Allocation*, Teknologi Informasi, Skill Gap

ABSTRACT

The dynamics of the information technology (IT) industry in Indonesia demand a massive adaptation of technical skill qualifications. The inability of job seekers and educational institutions to meet the real needs of this industry ultimately creates a significant skill gap. Previous efforts to map industry needs often used a supervised learning approach; however, this method tends to be static and rigid in dealing with the emergence patterns of new technical skills due to its reliance on labeled data. In response to these limitations, this study raises the research problem of how to build a system capable of identifying, evaluating, and mapping technical skill requirement patterns automatically. Through the extraction of job vacancy documents, this research aims to produce a model that can map the pillars of skill specialization that significantly dominate the job market.

This research applies a descriptive quantitative Natural Language Processing (NLP) method using the CRISP-DM framework. The data collection process was carried out through web scraping techniques on the LinkedIn portal for the Indonesian region during the fourth quarter of 2025. Prioritizing the ethics of public data retrieval, the extraction was executed automatically using a login-free scheme supported by the Selenium and BeautifulSoup libraries, successfully collecting 8,507 rows of raw data. Furthermore, the data went through a comprehensive preprocessing pipeline—including deduplication, data cleaning, case folding, normalization, compound terms handling, tokenization, document filtering, and whitelist-based filtering to extract pure hard skills. After the clean text was converted into a Bag-of-Words (BoW) matrix, an unsupervised learning algorithm in the form of Latent Dirichlet Allocation (LDA) was applied to map skill clusters by uncovering hidden semantic structures from the probability of word co-occurrences.

The Latent Dirichlet Allocation (LDA) algorithm proved effective in mapping IT skill specializations in Indonesia. Model evaluation demonstrated the best performance at a configuration of five topics (K=5), with a Coherence Score of 0.5637 (indicating strong semantic cohesion) and a Topic Diversity of 0.736 (representing clear cluster separation without overlapping). This research successfully answered the main problem by extracting five dominant skill pillars: Frontend & UI/UX Development, Cloud & DevOps Engineering, Data Science & Machine Learning, Backend Development, and Mobile Development, in which Frontend ranked at the top with a demand proportion of 36.62%. The output of this research contributes as an empirical foundation for higher education institutions to evaluate learning curricula, as well as a strategic compass for job seekers in aligning their skills with the actual qualifications absorbed by the technology industry.

Keywords: *Job Vacancies, Latent Dirichlet Allocation, Skill Gap, Skill Requirements, Topic Modeling*