BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi mendorong perusahaan untuk memanfaatkan sistem berbasis kecerdasan buatan dalam berbagai proses bisnis, termasuk pada bidang rekrutmen tenaga kerja [1], [2], [3]. Salah satu tahapan penting dalam rekrutmen adalah proses penyaringan *Curriculum Vitae (CV)* pelamar berdasarkan kesesuaian dengan deskripsi pekerjaan atau *job description (JD)* [4]. Pada proses konvensional, rekruter perlu membaca dan membandingkan setiap CV dengan kebutuhan posisi yang tersedia. Proses tersebut membutuhkan waktu yang cukup besar, terutama apabila jumlah pelamar banyak dan posisi yang dibuka memiliki kualifikasi teknis yang beragam [5].

Permasalahan pencocokan CV dan JD tidak hanya berkaitan dengan banyaknya dokumen yang harus diperiksa, tetapi juga berkaitan dengan perbedaan struktur dan gaya penulisan dokumen. CV dapat berisi ringkasan profil, pengalaman kerja, riwayat pendidikan, proyek, sertifikasi, dan daftar keterampilan [2]. Sementara itu, JD berisi uraian tanggung jawab pekerjaan, kualifikasi, pengalaman yang dibutuhkan, serta kemampuan teknis yang diharapkan. Perbedaan bentuk penyajian informasi ini menyebabkan pencocokan CV dan JD tidak selalu dapat dilakukan secara sederhana melalui pencocokan kata kunci [3].

Pada domain pekerjaan teknologi informasi, proses pencocokan CV dan JD menjadi lebih kompleks karena beberapa posisi memiliki keterkaitan keterampilan yang saling beririsan [6], [7], [8]. Misalnya, posisi *Backend Developer, Software Engineer, Full Stack Developer, DevOps Engineer, dan Database Administrator* dapat sama-sama memuat istilah seperti *API, database, Git, Docker, deployment, server, cloud, dan framework* pengembangan aplikasi. Kondisi tersebut menyebabkan batas antara kandidat yang benar-benar cocok dan kandidat yang hanya memiliki sebagian kesesuaian menjadi tidak selalu tegas. Oleh karena itu, diperlukan pendekatan yang mampu memahami hubungan makna antara isi CV dan JD secara lebih kontekstual [8].

Pendekatan berbasis *Natural Language Processing (NLP)* dapat digunakan untuk membantu proses pencocokan dokumen teks. Salah satu model NLP yang

banyak digunakan adalah *Bidirectional Encoder Representations from Transformers (BERT)*, yang mampu memahami konteks kata berdasarkan hubungan dua arah dalam kalimat [9], [10]. Dalam penelitian ini, model yang digunakan adalah *Multilingual BERT (mBERT)*, yaitu varian BERT yang dapat memproses teks dalam berbagai bahasa [11]. Pemilihan mBERT relevan karena data CV dalam penelitian ini mencakup bahasa Indonesia dan bahasa Inggris, serta memuat istilah teknis yang umum digunakan dalam industri teknologi informasi.

Penelitian ini menggunakan pendekatan *cross-encoder,* yaitu metode yang memproses pasangan CV dan JD secara bersamaan dalam satu input model [12], [13]. Pendekatan ini digunakan agar model dapat mempelajari hubungan semantik antara dua dokumen secara langsung. Format masukan model disusun dalam bentuk pasangan teks, yaitu CV sebagai teks pertama dan JD sebagai teks kedua. Dari pasangan tersebut, model menghasilkan skor probabilitas yang menunjukkan tingkat kecocokan antara CV dan JD [10].

Selain klasifikasi kecocokan, penelitian ini juga memperhatikan aspek *ranking.* Dalam konteks rekrutmen, sistem tidak cukup hanya memberikan label cocok atau tidak cocok, tetapi juga perlu membantu menyusun urutan kandidat atau *job description* berdasarkan tingkat kecocokan tertinggi [14], [11]. Oleh karena itu, skor probabilitas yang dihasilkan model digunakan sebagai dasar untuk melakukan pemeringkatan. Dengan adanya mekanisme *ranking,* sistem diharapkan dapat membantu rekruter dalam memprioritaskan kandidat yang paling relevan terhadap suatu posisi [15].

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk mengimplementasikan model *Multilingual BERT* dalam klasifikasi dan *ranking* kecocokan CV pelamar kerja terhadap deskripsi pekerjaan pada domain teknologi informasi. Penelitian ini mengevaluasi performa model menggunakan metrik klasifikasi dan *ranking*, serta mengimplementasikan model terbaik ke dalam prototipe aplikasi berbasis Streamlit sebagai simulasi penggunaan sistem *screening* awal. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pemanfaatan NLP untuk mendukung proses rekrutmen secara lebih terstruktur, efisien, dan berbasis data.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas dapat dirumuskan permasalahan utama dalam penelitian ini sebagai berikut:

1. Bagaimana implementasi model *Multilingual BERT (mBERT)* dengan pendekatan *cross-encoder* untuk melakukan klasifikasi dan *ranking* kecocokan antara *Curriculum Vitae* dan deskripsi pekerjaan pada domain Teknologi Informasi?
2. Bagaimana performa model mBERT *cross-encoder* dalam melakukan klasifikasi dan *ranking* kecocokan *Curriculum Vitae* terhadap deskripsi pekerjaan berdasarkan metrik evaluasi yang digunakan?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah, terukur, dan sesuai dengan ruang lingkup yang telah ditetapkan, maka batasan masalah dalam penelitian ini dirumuskan sebagai berikut:

1. Penelitian ini hanya berfokus pada pencocokan dokumen *Curriculum Vitae (CV)* dan deskripsi pekerjaan berbasis teks, tanpa mempertimbangkan faktor non-teks seperti wawancara, psikotes, portofolio visual, referensi kerja, atau penilaian HR secara langsung.
2. Domain penelitian dibatasi pada bidang Teknologi Informasi (IT) dengan 10 *role family,* yaitu *Network / Infrastructure, Database / Infrastructure, Backend Developer, Frontend Developer, Web Developer, DevOps Engineer, Mobile Developer, AI / Data Science, Software Engineer, dan Full Stack Developer.*
3. *Dataset* CV akhir yang digunakan dalam eksperimen utama terdiri dari 500 CV, dengan komposisi 50 CV per *role family* dan pembagian bahasa yang seimbang, yaitu 25 CV berbahasa Inggris dan 25 CV berbahasa Indonesia pada setiap *role*.
4. *Dataset* deskripsi pekerjaan yang digunakan berjumlah 2.264 dokumen JD yang telah melalui proses pembersihan dan standardisasi *role family*.
5. Model yang digunakan terbatas pada Multilingual BERT (mBERT) dengan *checkpoint bert-base-multilingual-cased* dan pendekatan *cross-encoder.*

Penelitian ini tidak melakukan perbandingan dengan model *transformer* lain seperti IndoBERT, RoBERTa, XLM-R, atau *Sentence-BERT.*

6. Klasifikasi dilakukan secara biner dengan label cocok dan tidak cocok. Kategori interpretatif seperti Cocok, Cukup Cocok, dan Tidak Cocok ditentukan berdasarkan skor probabilitas dan *threshold* yang ditetapkan.
7. *Pairing* data dilakukan menggunakan pendekatan *skill-aware pairing*, dengan 10 pasangan positif dan 15 pasangan negatif untuk setiap CV. *Negative sampling* dibentuk dari pasangan antar-role IT yang berbeda, bukan dari kategori *Non-IT*.
8. Proses tokenisasi dibatasi pada panjang maksimum 512 token. Bagian teks yang melebihi batas tersebut akan dipotong melalui mekanisme *truncation*.
9. Evaluasi model dilakukan secara kuantitatif menggunakan metrik klasifikasi dan *ranking*, tanpa melibatkan validasi langsung oleh praktisi HR.
10. *Deployment* dalam penelitian ini dibatasi pada prototipe aplikasi berbasis Streamlit sebagai simulasi implementasi model. Aplikasi yang dibangun bukan merupakan sistem *Applicant Tracking System* (ATS) produksi dan belum terintegrasi dengan sistem rekrutmen perusahaan secara langsung.
11. Penelitian ini hanya menggunakan teks CV dan teks *job description* sebagai input model, bukan dokumen utuh dalam format PDF, DOC, atau DOCX. Informasi visual dan struktur dokumen, seperti tata letak, tabel, foto, ikon, serta format penulisan, tidak diproses oleh model.

1.4 Tujuan

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan model Multilingual BERT dengan pendekatan *cross-encoder* untuk melakukan klasifikasi kecocokan CV-JD serta menyusun *ranking* berdasarkan skor probabilitas kelas Cocok.
2. Mengevaluasi performa model dalam klasifikasi dan *ranking* menggunakan metrik klasifikasi dan metrik *ranking*

.

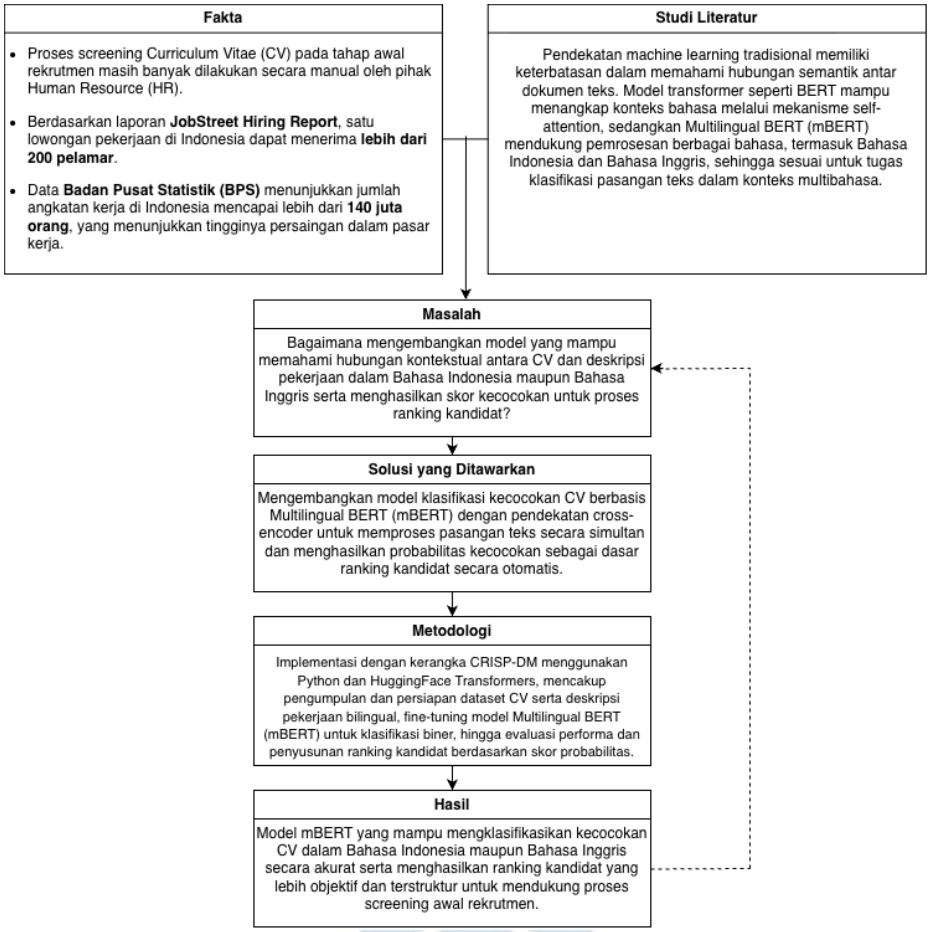
1.5 Manfaat

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Teoretis: Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan kajian *Natural Language Processing* (NLP), khususnya pada penerapan arsitektur *transformer* Multilingual BERT (mBERT) untuk tugas klasifikasi pasangan teks (*text pair classification*). Selain itu, penelitian ini dapat memperkaya referensi akademik terkait pemodelan hubungan semantik antara *Curriculum Vitae* (CV) dan deskripsi pekerjaan menggunakan pendekatan *cross-encoder,* serta menjadi dasar bagi penelitian lanjutan dalam pengembangan sistem berbasis kecerdasan buatan pada bidang rekrutmen dan seleksi tenaga kerja.
2. Manfaat Praktis: Secara praktis, penelitian ini diharapkan dapat membantu proses *screening* awal pelamar kerja dengan menyediakan model klasifikasi kecocokan CV yang mampu menghasilkan skor probabilitas sebagai dasar *ranking* kandidat secara objektif dan terstruktur. Hasil penelitian ini dapat dimanfaatkan sebagai referensi pengembangan sistem pendukung keputusan dalam proses rekrutmen, sehingga dapat meningkatkan efisiensi waktu seleksi dan mengurangi beban kerja pada tahap penyaringan awal dokumen pelamar.
3. Manfaat Metodologis Studi: Penelitian ini diharapkan dapat memberikan gambaran tahapan pengembangan model klasifikasi berbasis *transformer* menggunakan pendekatan *cross-encoder* pada pasangan teks, mulai dari proses persiapan *dataset,* pelabelan data, preprocessing, *partial fine-tuning* model Multilingual BERT (mBERT), hingga evaluasi performa dan mekanisme *ranking* berbasis probabilitas. Alur metodologi yang disusun dalam penelitian ini dapat menjadi referensi dalam pengembangan penelitian serupa yang memanfaatkan model pretrained *transformer* untuk tugas klasifikasi dokumen atau pencocokan teks pada konteks lainnya.

1.6 Kerangka Berpikir

Kerangka berpikir penelitian ini dirancang untuk memetakan alur logis penelitian secara sistematis, mulai dari identifikasi konteks yang lebih luas hingga luaran spesifik yang diharapkan, sebagaimana divisualisasikan pada Gambar 1.1.



Gambar 1. 1 Kerangka Berpikir

Proses *screening Curriculum Vitae* (CV) dalam tahap awal rekrutmen umumnya masih dilakukan secara manual, sehingga memerlukan waktu yang cukup lama dan berpotensi menimbulkan subjektivitas dalam penilaian. Permasalahan ini semakin kompleks ketika dokumen CV dan deskripsi pekerjaan ditulis dalam Bahasa Indonesia maupun Bahasa Inggris, sehingga proses pencocokan memerlukan pemahaman konteks bahasa yang baik. Kondisi tersebut menunjukkan perlunya solusi berbasis teknologi yang mampu membantu proses klasifikasi dan penentuan kecocokan kandidat secara lebih efisien dan objektif. Berdasarkan studi literatur, pendekatan *machine learning* tradisional memiliki keterbatasan dalam memahami hubungan semantik antar dokumen teks. Perkembangan arsitektur *transformer* seperti BERT memungkinkan pemodelan konteks bahasa yang lebih mendalam melalui mekanisme *self-attention*. Multilingual BERT (mBERT) secara khusus dirancang untuk mendukung

pemrosesan berbagai bahasa, termasuk Bahasa Indonesia dan Bahasa Inggris, sehingga relevan untuk digunakan dalam kasus *screening* CV berbasis teks multibahasa.

Berdasarkan fakta dan kajian literatur tersebut, permasalahan penelitian ini difokuskan pada bagaimana mengembangkan model yang mampu memahami hubungan kontekstual antara CV dan deskripsi pekerjaan dalam dua bahasa serta menghasilkan skor kecocokan yang dapat digunakan untuk melakukan *ranking* kandidat secara otomatis.

Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan pengembangan model klasifikasi berbasis Multilingual BERT (mBERT) dengan pendekatan *cross-encoder*, di mana pasangan teks CV dan deskripsi pekerjaan diproses secara simultan untuk menghasilkan probabilitas kecocokan. Penelitian ini dilaksanakan menggunakan kerangka CRISP-DM yang meliputi tahapan pemahaman masalah, persiapan data, pemodelan, dan evaluasi. Model yang dikembangkan akan dilatih untuk melakukan klasifikasi biner (cocok dan tidak cocok), dan skor probabilitas yang dihasilkan akan digunakan sebagai dasar penyusunan *ranking* kandidat. Hasil yang diharapkan dari penelitian ini adalah diperolehnya model mBERT yang mampu mengklasifikasikan kecocokan CV secara akurat dalam Bahasa Indonesia maupun Bahasa Inggris serta menghasilkan sistem *ranking* kandidat yang lebih objektif dan terstruktur dalam mendukung proses *screening* awal rekrutmen.

1.7 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum tentang proses penulisan yang dilakukan untuk menyelesaikan penelitian ini. Sistematika penulisan penelitian ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi uraian mengenai latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, kerangka berpikir, dan sistematika penulisan. Bagian ini menjelaskan dasar pemikiran dilakukannya penelitian, permasalahan yang ingin diselesaikan, serta arah penelitian yang digunakan sebagai landasan untuk bab-bab berikutnya.

BAB II TINJAUAN PUSTAKA

Bab ini memuat teori dan konsep yang digunakan sebagai dasar dalam penelitian. Pembahasan pada bab ini mencakup konsep *Curriculum Vitae*, *job description*, rekrutmen, *Natural Language Processing*, *text classification*, *text matching*, BERT, Multilingual BERT, *cross-encoder*, metrik evaluasi klasifikasi, metrik evaluasi *ranking,* serta penelitian terdahulu yang relevan. Landasan teori digunakan untuk memperkuat dasar ilmiah dari metode yang diterapkan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan penelitian yang dilakukan dengan mengacu pada metode CRISP-DM. Pembahasan meliputi tahap *business understanding*, *data understanding*, *data preparation*, modeling, *evaluation*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari setiap tahapan penelitian, mulai dari hasil pemahaman bisnis, karakteristik data, proses persiapan data, kronologi eksperimen, hasil pelatihan model, evaluasi model pada data uji, analisis *threshold,* mekanisme *ranking,* analisis kesalahan, hingga hasil *deployment* aplikasi Streamlit. Bab ini juga memuat pembahasan mengenai peningkatan performa model dari *baseline* sampai eksperimen final, faktor yang memengaruhi performa model, serta interpretasi hasil klasifikasi dan *ranking*.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan berdasarkan rumusan masalah dan tujuan penelitian. Selain itu, bab ini juga memuat saran untuk pengembangan penelitian selanjutnya.