BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang semakin maju, data teks telah menjadi salah satu aset berharga. Dari dokumen ilmiah hingga postingan media sosial, teks hadir dalam jumlah yang sangat besar dan beragam [1]. Teks merupakan komponen utama dari informasi yang banyak digunakan saat ini. Menurut laporan dari Republika.co.id, jumlah data digital yang dihasilkan setiap hari mencapai lebih dari 2,5 eksabita, dengan sebagian besar berasal dari teks yang dihasilkan pengguna [2]. Oleh karena itu, kemampuan yang efisien dan efektif dalam menganalisis teks ini menjadi semakin penting.

Pengelolaan dan analisis data teks menghadirkan tantangan yang kompleks. Salah satu tantangan utama adalah deteksi kemiripan, yaitu kemampuan untuk menentukan seberapa mirip dua teks atau lebih. Metode konvensional seperti penghitungan *cosine*, *jaccard*, dan pendekatan berbasis *bag-of-words* telah digunakan secara luas [3]. *Cosine similarity*, salah satu metode yang paling umum digunakan dengan cara menghitung sudut kosinus antara dua vektor teks yang mewakili dua dokumen. Meskipun metode ini cukup efektif dalam beberapa kasus, metode tersebut memiliki keterbatasan yang signifikan. Misalnya, metode ini sering kali gagal menangkap makna kontekstual dari kata, sehingga sulit untuk mengidentifikasi sinonim atau frasa yang memiliki arti yang sama tetapi ditulis dengan cara yang berbeda [4].

Untuk mengatasi keterbatasan metode konvensional, pengembangan pada *Natural Language Processing* (NLP) telah memperkenalkan algoritma yang lebih canggih dan kontekstual. Salah satu inovasi terbesar dalam dekade terakhir adalah *Bidirectional Encoder Representations from Transformers* (BERT) [5]. *Google* memperkenalkan algoritma BERT pada tahun 2018, algoritma ini dirancang untuk memahami konteks dua arah dalam teks, berbeda dengan model sebelumnya yang hanya memproses teks dari kiri ke kanan atau sebaliknya. Algoritma *BERT* dapat memahami makna yang lebih dalam dan lebih akurat dari teks, hal ini membuatnya sangat efektif dalam tugas-tugas seperti deteksi kemiripan [6].

Meskipun *BERT* secara signifikan meningkatkan akurasi dalam mendeteksi kemiripan, penerapannya dalam sistem ini masih menghadapi beberapa tantangan. Pertama, model *BERT* sangat kompleks dan besar, yang berarti membutuhkan sumber daya komputasi yang tinggi untuk melatih dan menjalankannya [7]. Kedua, integrasi *BERT* ke dalam sistem yang sudah ada memerlukan pengetahuan teknis yang mendalam dan penyesuaian yang cermat untuk memastikan bahwa model berjalan efisien dan efektif. Ketiga, interpretasi hasil yang dihasilkan oleh *BERT* bisa menjadi sulit karena kompleksitas model, yang kadang menghasilkan hasil yang ambigu atau tidak konsisten jika tidak diimplementasikan dengan benar [8]. Oleh karena itu, diperlukan metode alternatif yang dapat diintegrasikan bersama *BERT* untuk meningkatkan kinerja deteksi kemiripan.

Penelitian sebelumnya pada deteksi kemiripan umumnya hanya berfokus pada satu metode, seperti penggunaan algoritma *Cosine similarity*. Seperti penelitian yang dilakukan oleh Safriandi Hatoguan Sihombing, dkk yang berjudul "*Implementasi Metode Cosine similarity untuk Deteksi Kemiripan Data pada Dokumen Word dan Perbedaan Isi Dokumen Word*", penelitian ini lebih fokus pada kesamaan keseluruhan teks tanpa mempertimbangkan bagaimana sinonim dan antonim dapat mengubah makna dalam konteks yang lebih luas [9]. Selain itu, sebagian besar penelitian tentang deteksi kemiripan belum banyak mengeksplorasi penerapan gabungan algoritma untuk meningkatkan akurasi analisis teks [10].

Penelitian ini bertujuan untuk mengembangkan perangkat bantu deteksi kemiripan teks yang mengintegrasikan algoritma *Cosine similarity* dengan *BERT* untuk mendeteksi kemiripan baik dari segi sintaksis maupun semantik secara lebih menyeluruh. Dengan pendekatan tersebut, penelitian ini diharapkan dapat meningkatkan akurasi deteksi kemiripan lintas bahasa, memberikan kontribusi yang signifikan terhadap pemahaman dan aplikasi algoritma dalam konteks multi-bahasa. Dengan tujuan tersebut, diharapkan perangkat bantu yang akan dikembangkan dapat memberikan kontribusi yang signifikan dalam analisis teks. Maka dari itu, tugas akhir ini berjudul "**Perangkat Bantu Deteksi *Similarity* Menggunakan Algoritma *BERT* dan *Cosine similarity***."

1.2 Rumusan Masalah

Berdasarkan latar belakang permasalahan di atas maka dapat dirumuskan permasalahan penelitian, sebagai berikut:

1. Bagaimana membangun aplikasi yang mampu mengidentifikasi potensi kemiripan teks dalam berbagai bahasa secara efektif?
2. Bagaimana penerapan algoritma BERT dan Cosine Similarity dapat digunakan untuk mengukur tingkat kemiripan teks secara akurat**?**
3. Seberapa besar nilai *Mean Absolute Error* (*MAE*) dari hasil deteksi kemiripan naskah?

1.3 Tujuan Penelitian

Tujuan penelitian ini, sebagai berikut:

1. Membangun aplikasi yang mampu mengidentifikasi potensi kemiripan multi-bahasa.
2. Menerapkan algoritma BERT dan Cosine Similarity pada aplikasi untuk menyajikan tingkat kemiripan naskah.
3. Mengetahui nilai *Mean Absolute Error* (*MAE*) dari hasil deteksi kemiripan naskah

1.4 Batasan Masalah

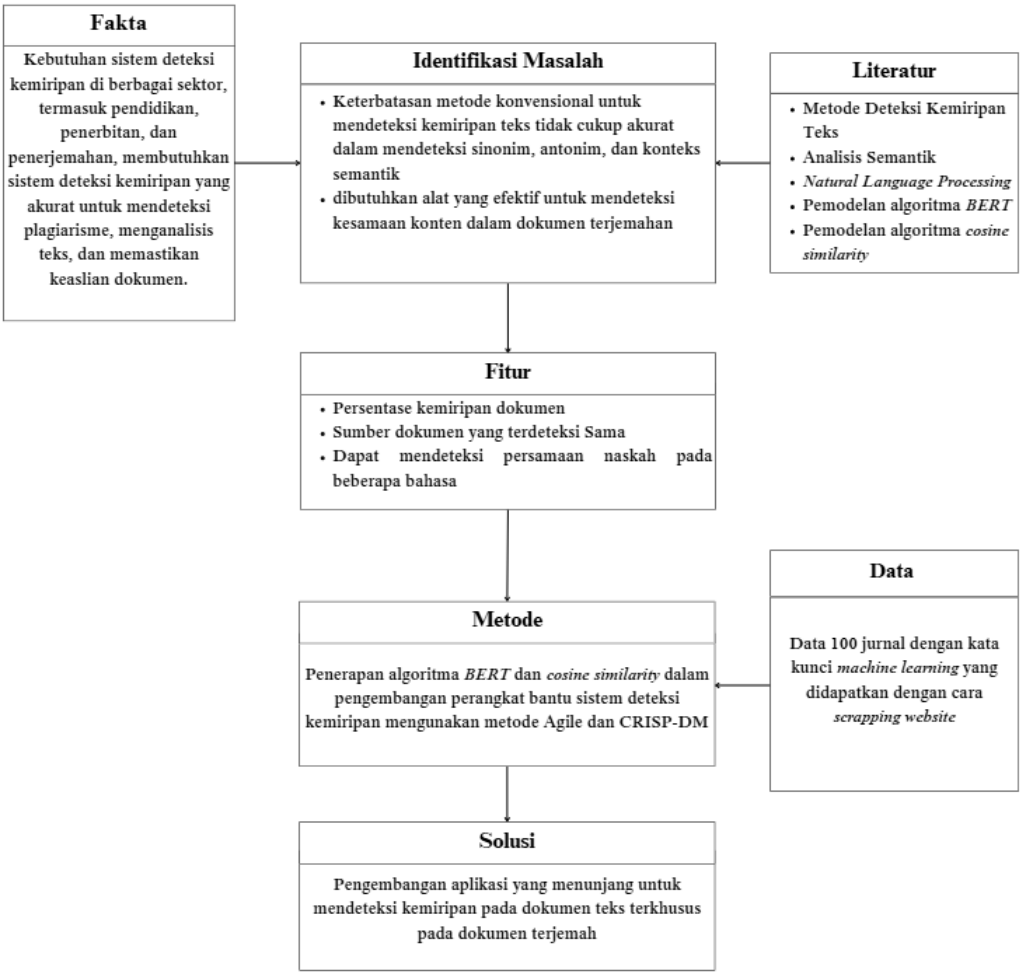
Batasan masalah dalam penelitian ini dirumuskan untuk memberikan fokus pada aspek-aspek tertentu yang relevan dengan pengembangan sistem deteksi kemiripan menggunakan algoritma *BERT*. Berikut adalah batasan masalah yang telah ditetapkan:

1. Penelitian ini hanya akan fokus pada integrasi algoritma *BERT* dan *cosine similarity* ke dalam sistem deteksi kemiripan pada teks.
2. Analisis semantik yang akan ditambahkan terbatas pada sinonim dan antonim tidak akan mencakup analisis semantik menggunakan pendekatan lain.
3. Deteksi kesamaan konten terjemahan akan difokuskan pada Indonesia, Prancis, German, Inggris, Jepang, dan Rusia.
4. Dataset yang digunakan untuk pengujian berbahasa Indonesia kurang lebih 100 jurnal yang bertema "*Machine Learning"* bersumber dari *Google Scholar*.

5. Penelitian ini menggunakan dua metode, yaitu Agile dan CRISP-DM, dengan batasan yang telah ditentukan untuk masing-masing metode. Pada metode Agile, *step* yang digunakan meliputi *planning*, desain, *development, testing*, dan *review*, yang berfokus pada pendekatan iteratif dan fleksibel dalam pengembangan aplikasi. Sementara itu, metode CRISP-DM digunakan untuk mendukung proses analisis data, dengan *step* yang digunakan mencakup data *understanding*, data *preparation*, dan *modelling.*

1.5 Kerangka Pemikiran

Kerangka pemikiran adalah suatu proses di mana penelitian merencanakan cara untuk menyusun pertanyaan-pertanyaan yang akan diajukan dan mendorong penyelidikan atas masalah-masalah yang ada, serta menyediakan latar belakang dan konteks yang menjelaskan mengapa penelitian tersebut dilakukan [11]. Adapun kerangka pemikiran dari penelitian ini terdapat pada gambar 1.1.



Gambar 1.1 Kerangka Pemikiran

Gambar 1.1 di atas merupakan kerangka pemikiran penelitian yang menggambarkan alur logis dari permasalahan hingga solusi dalam pengembangan sistem deteksi kemiripan dokumen berbasis algoritma *BERT* dan *cosine similarity*. Kerangka penelitian ini dimulai dengan fakta yang menunjukan bahwa sistem deteksi kemiripan teks menjadi kebutuhan penting di berbagai bidang, seperti pendidikan, penerbitan, dan penerjemahan, karena banyak digunakan untuk mendeteksi plagiarisme, menganalisis teks, serta memastikan keaslian dokumen. Namun, metode konvensional seperti *cosine similarity* dan *jaccard* hanya mampu mendeteksi kesamaan teks secara sintaksis tanpa menangkap makna semantik yang lebih mendalam, sehingga seringkali hasilnya tidak akurat. Oleh karena itu, diperlukan alat yang lebih efektif yang tidak hanya mempertimbangkan kesamaan kata, tetapi juga konteks dan sinonim dari teks.

Penelitian ini juga menggabungkan algoritma *BERT* dan *cosine similarity* untuk menghasilkan sistem deteksi kemiripan yang lebih akurat. Fitur yang dikembangkan meliputi kemampuan menghitung persentase kemiripan dokumen, menampilkan sumber teks yang mirip, dan mendeteksi kesamaan teks lintas bahasa. Metode yang digunakan dalam penelitian ini adalah pendekatan CRISP-DM yang mencakup tahapan *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment*. Data yang digunakan berasal dari 100 jurnal ilmiah dengan kata kunci "*Machine Learning"* yang diperoleh melalui *scraping website*, dan digunakan untuk melatih serta menguji model. Hasil dari penelitian ini adalah sebuah aplikasi yang mampu mendeteksi kemiripan teks dengan lebih akurat dibandingkan metode konvensional, sehingga diharapkan dapat membantu dalam mendeteksi plagiarisme dan menganalisis kemiripan teks lintas bahasa secara lebih efektif dan efisien.

1.6 Sistematika Penulisan

Sistematika penulisan setiap bab dalam laporan tugas akhir mempunyai tujuan yang berbeda sehingga dapat memudahkan bagi pembaca untuk dipahami dan dimengerti. Pada penelitian ini, penulis membagi sistematika penulisan menjadi lima bab diantaranya:

BAB 1: PENDAHULUAN

Bab ini menjelaskan latar belakang penelitian, rumusan masalah, tujuan penelitian, batasan masalah, kerangka pemikiran, dan menjelaskan cara penulisan secara sistematis.

BAB II: TINJAUAN LITERATUR

Bab ini membahas tentang perkembangan paling mutakhir dalam dunia keilmuan atau sering disebut dengan *state of the art* dari teori yang sedang dikaji dan kedudukan masalah penelitian dalam bidang informatika yang diteliti

BAB III: METODOLOGI PENELITIAN

Bab ini menguraikan metode yang digunakan dalam penelitian. Selain itu, bab ini merumuskan tahapan apa saja yang akan dilakukan selama penelitian. Menjelaskan langkah-langkah dan teknik yang dilakukan dalam penelitian, dijelaskan secara kronologis dan sistematis.

BAB IV: HASIL DAN PEMBAHASAN

Bab ini berisi hasil implementasi dan pengujian. Pada bab ini dipaparkan dua hal utama, pertama pemaparan tentang temuan atau hasil penelitian berdasarkan tahapan penelitian yang dilakukan. Kedua pembahasan hasil atau temuan penelitian untuk menjawab rumusan penelitian.

BAB V: SIMPULAN DAN SARAN

Bab ini menguraikan kesimpulan dari pembahasan yang dilakukan dan membahas serta memberikan saran bagi peneliti selanjutnya Penulisan simpulan disampaikan dengan cara uraian padat lebih baik daripada dengan cara butir demi butir.

UNIVERSITAS ISLAM NEGERI
BANDUNG