BAB I
PENDAHULUAN

1.1 Latar Belakang

Deteksi plagiarisme, khususnya plagiarisme parafrastik, merupakan salah satu permasalahan mendasar dalam pengolahan bahasa alami (*Natural Language Processing*), karena dua teks dapat menyampaikan gagasan yang sama meskipun disajikan dengan struktur dan pilihan kata berbeda. Pemeriksaan manual memakan waktu lama dan rawan subjektivitas, sehingga dibutuhkan pendekatan komputasional yang efisien [1]. Tantangan ini kian berat selain karena bisa membuat kemiripan leksikal menjadi 0 tapi makna tetap sama, plagiarisme kini semakin banyak disamarkan dengan parafrase otomatis oleh *Large Language Model* (LLM) seperti *ChatGPT* dan juga kasus plagiarisme lintas bahasa (terjemahan lalu parafrase) sehingga struktur leksikal bisa hilang dan sulit ditangkap oleh metode konvensional seperti *string-matching* (membandingkan urutan karakter secara langsung) [2]. Parafrase modern umumnya memanfaatkan mekanisme, seperti substitusi sinonim, pengubahan urutan kata atau frasa (*reordering*) serta penyisipan-penghapusan yang secara sistematis melemahkan jejak leksikal [3]. Sejalan dengan fenomena tersebut, *Global Plagiarism Report* 2018–2024 mencatat bahwa tingkat plagiarisme global mengalami peningkatan hingga lebih dari 20% pada tahun 2023, yang salah satunya dipengaruhi oleh semakin luasnya penggunaan teknologi penulisan berbasis AI [4]. Oleh karena itu, *Semantic Textual Similarity* yang mengukur derajat kesamaan makna antarteks, menjadi kerangka yang lebih tepat dibanding pendekatan leksikal semata karena menilai kesepadanan makna [5]. *Semantic Textual Similarity* (STS) juga sudah menjadi kerangka evaluasi tahunan bersama di komunitas riset misalnya melalui program *SemEval*, menekankan urgensinya dalam dunia *Natural Language Processing* [6].

Pendekatan berbasis *Deep Learning* telah digunakan dalam *Semantic Textual Similarity* (STS) melalui model *Recurrent Neural Network* (RNN). Algoritma tersebut dirancang untuk memproses *input* berurutan (sequential) seperti pada data teks dan menyimpan konteks serta memori antartoken (antar kata) [7]. Meskipun demikian, *Recurrent Neural Network* (RNN) konvensional memiliki kelemahan terkait dengan *vanishing gradient*, yaitu kondisi ketika urutan teks terlalu panjang sehingga informasi pada bagian awal cenderung terabaikan, yang pada akhirnya mengurangi kemampuan model dalam menangkap konteks jangka panjang [8,9]. Salah satu pengembangan dari *Recurrent Neural Network* (RNN) untuk mengatasi permasalahan tersebut adalah model *Long Short-Term Memory* (LSTM), yang dilengkapi

dengan mekanisme gerbang (*gate*) untuk mengatur dan mempertahankan informasi penting secara lebih efektif [10]. LSTM memiliki beberapa gerbang seperti *forget gate*, *input gate*, dan *output gate* yang memungkinkan model menyimpan informasi relevan dalam jangka waktu lebih panjang.

Sebagai penguatan lebih lanjut, LSTM yang memproses urutan teks dari dua arah, yaitu maju dan mundur, atau dikenal sebagai *Bidirectional* LSTM (BiLSTM), terbukti mampu menangkap konteks secara lebih utuh karena mempertimbangkan hubungan antarkata baik dari konteks sebelumnya maupun sesudahnya [11,12]. Keunggulan BiLSTM terletak pada kemampuannya dalam memahami perubahan makna kata berdasarkan urutan kemunculannya dalam kalimat, termasuk fenomena linguistik seperti negasi, penekanan, dan pergeseran makna semantik, di mana arti suatu kata atau frasa dapat berubah bergantung pada kata yang mendahului maupun mengikutinya [13]. Dengan demikian, representasi semantik yang dihasilkan tidak hanya bergantung pada kata-kata terdekat, tetapi juga pada struktur dan susunan kalimat secara keseluruhan.

Selain itu, penerapan mekanisme *additive attention* di atas lapisan BiLSTM memungkinkan model untuk memfokuskan perhatian pada bagian teks yang paling informatif, terutama ketika berhadapan dengan *input* yang lebih panjang [14]. Mekanisme ini membantu model menyaring informasi yang kurang relevan dan mengekstrak intisari makna kalimat, sehingga representasi akhir yang dihasilkan lebih mencerminkan bagian teks yang benar-benar berkontribusi terhadap kesamaan makna antarteks [15]. BiLSTM dan *attention* dalam beberapa tahun terakhir mulai semakin banyak digunakan dalam dunia *Natural Langunage Processing* secara umum [16, 17].

Berdasarkan uraian di atas, penelitian ini mengembangkan sebuah aplikasi yang mampu mengukur tingkat kesamaan makna antarteks (*Semantic Textual Similarity*) sebagai upaya untuk mendeteksi indikasi plagiarisme. Sistem ini diimplementasikan menggunakan *Bidirectional Long Short-Term Memory* (BiLSTM) dan mekanisme *attention* dalam platform berbasis *web*, dengan judul tugas akhir "**Implementasi BiLSTM dan Attention pada Aplikasi Semantic Textual Similarity Berbasis Web**".

1.2 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah tugas akhir ini adalah sulitnya mendeteksi plagiarisme parafrastik karena teks yang memiliki makna sama dapat disajikan dengan struktur dan pilihan kata yang berbeda, sehingga metode berbasis kesamaan leksikal sulit menangkap kesamaan makna antarteks.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah mengembangkan dan mengimplementasikan aplikasi *web* untuk mengukur kesamaan semantik antarteks guna mendeteksi plagiarisme dengan memanfaatkan metode BiLSTM dengan mekanisme *attention*.

1.4 Manfaat

Manfaat dari tugas akhir ini adalah:

1. Menyediakan aplikasi berbasis *web* yang dapat membantu mendeteksi indikasi plagiarisme parafrastik dan mengurangi ketergantungan pada pemeriksaan manual.
2. Menjadi referensi dalam pengembangan penelitian dibidang *Natural Language Processing*, khususnya pada pengukuran kesamaan semantik dan deteksi plagiarisme berbasis *Deep Learning*.

1.5 Ruang Lingkup

Ruang lingkup dari tugas akhir ini mencakup:

1. Aplikasi *web* yang dikembangkan hanya berfungsi untuk membandingkan dua *input* teks yang diberikan pengguna, baik berupa perbandingan antarkalimat maupun antardokumen. Hasil keluaran sistem berupa skor kesamaan semantik dalam bentuk persentase antara 0 hingga 100.
2. *Dataset* yang digunakan dalam penelitian ini adalah *Quora Question Pairs* (QQP) yang diperoleh dari (*https://gluebenchmark.com/tasks/*) pada tahun 2017, dengan jumlah data sekitar 400.000 pasangan pertanyaan. *Dataset* ini dipilih karena sangat banyak digunakan sebagai benchmark dalam penelitian *Semantic Textual Similarity* (STS) serta memiliki skala data yang jauh lebih besar dibandingkan dengan *dataset* STS lainnya.
3. Seluruh proses pelatihan dan pengujian dilakukan menggunakan bahasa Inggris, yang merupakan bahasa asli pada *dataset* QQP. Oleh karena itu, penelitian ini tidak mencakup teks dalam bahasa lain dan dibatasi pada analisis kesamaan semantik teks berbahasa Inggris.
4. Tahap pelatihan model (*training*) dilakukan dengan membandingkan pasangan kalimat dari *dataset* yang telah ditentukan. Model tidak dilatih menggunakan paragraf panjang maupun kumpulan dokumen.
5. Proses pengujian (*testing)* pada level dokumen menggunakan model yang telah dilatih pada pasangan kalimat dari *dataset*. Setiap dokumen akan dipecah menjadi beberapa bagian

(segmen kalimat) dan dilakukan perbandingan secara berpasangan. Skor akhir kesamaan dokumen diperoleh dari rata-rata 3 skor tertinggi dari pasangan antar segmen.

6. Arsitektur yang digunakan dalam penelitian ini mengikuti pendekatan *cross-encoder*, di mana dua kalimat digabungkan sebagai satu input tunggal dan diproses secara bersama oleh satu *encoder* BERT. Representasi keluaran *encoder* tersebut kemudian diproses lebih lanjut menggunakan satu lapisan BiLSTM yang kemudian diikuti dengan mekanisme *Attention*. Setiap kalimat tidak memiliki *encoder* BERT maupun BiLSTM yang terpisah.
7. BERT digunakan sebagai *feature extractor* (*frozen*) tanpa dilakukan *fine-tuning*, dan hanya berfungsi untuk menghasilkan *embedding* dari teks masukan.
8. *Input* teks antarkalimat pada aplikasi *web* dibatasi maksimal 30 kata untuk setiap kalimat. Apabila jumlah kata melebihi batas tersebut, sistem akan melakukan pemotongan otomatis (*truncation*). Batasan ini ditetapkan berdasarkan karakteristik *dataset Quora Question Pairs*, di mana sebagian besar pasangan pertanyaan terdiri dari teks pendek, sehingga pembatasan panjang *input* diperlukan agar distribusi data pada tahap pengujian tetap sebanding dengan data pelatihan.
9. Untuk pengujian pada tingkat dokumen, panjang dokumen dibatasi hingga maksimal dua halaman, atau sekitar 1.000 kata maksimal, dengan pertimbangan keterbatasan sumber daya komputasi pada sistem yang digunakan.
10. Format dokumen yang dapat digunakan sebagai masukan dalam perbandingan antardokumen terbatas pada *file* berformat DOCX (*Word*), PDF, dan TXT, dengan panjang teks maksimal yang telah ditentukan.