BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa alami secara otomatis [1, 2]. Perkembangan NLP dalam beberapa tahun terakhir didorong oleh kemajuan dalam model pembelajaran mendalam berbasis Transformer, yang mampu memahami konteks bahasa secara lebih akurat dan efisien [2, 3]. NLP kini banyak diterapkan dalam berbagai bidang seperti analisis sentimen [4, 5], deteksi ujaran kebencian [6, 7], dan deteksi teks buatan AI [8, 9], dengan peningkatan performa yang signifikan berkat penggunaan model bahasa besar berbasis arsitektur Transformer [3].

Model bahasa besar (Large Language Models/LLMs) seperti GPT-3 dan GPT-4 telah menunjukkan kemampuan luar biasa dalam menghasilkan teks yang koheren dan kontekstual menyerupai tulisan manusia [10, 11]. Kemampuan ini menjadikan LLMs banyak dimanfaatkan dalam berbagai aplikasi, seperti chatbot, penulisan otomatis, hingga asisten virtual [12, 13]. Salah satu contohnya adalah ChatGPT, yang berhasil menarik perhatian global dengan kemampuannya dalam menghasilkan teks menyerupai tulisan manusia dan berinteraksi secara kontekstual. Dalam waktu hanya dua bulan sejak peluncurannya, ChatGPT mencapai lebih dari 100 juta pengguna [14]. Namun, kemajuan ini juga menimbulkan kekhawatiran terhadap potensi penyalahgunaan, seperti penyebaran misinformasi [15, 16], penipuan [17], dan pelanggaran integritas akademik [18]. Oleh karena itu, penting untuk mengembangkan metode deteksi teks buatan AI secara otomatis guna membedakan teks yang ditulis oleh manusia dengan yang dihasilkan oleh model generatif seperti ChatGPT [8, 19].

Salah satu pendekatan yang banyak digunakan dalam mendeteksi teks buatan adalah melalui pemanfaatan model pembelajaran mesin dan pembelajaran mendalam, khususnya model berbasis arsitektur Transformer yang telah terbukti efektif dalam memahami pola-pola bahasa yang kompleks [15, 20]. Model deteksi ini sering kali dilatih pada korpus yang terdiri dari kombinasi teks buatan dan teks asli, sehingga memungkinkan sistem untuk mempelajari perbedaan distribusi linguistik antara keduanya [8]. Selain itu, teknik berbasis embedding semantik,

klasifikasi berbasis fine-tuning model pralatih, serta penggunaan discriminator sebagai pendeteksi juga telah diterapkan dengan tingkat akurasi yang cukup tinggi [15, 21].

Beberapa studi terdahulu telah menunjukkan hasil yang menjanjikan dalam mendeteksi teks buatan. Sebuah penelitian menunjukkan bahwa penggunaan detektor berbasis GPT-2 mampu mengidentifikasi teks buatan secara lebih akurat dibandingkan penilaian manusia, meskipun jumlah data yang digunakan masih terbatas [22]. Studi lain menunjukkan bahwa teks yang dihasilkan oleh ChatGPT cenderung memiliki lebih sedikit kesalahan tata bahasa dan lebih banyak token unik dibandingkan dengan teks manusia, yang mengindikasikan adanya pola linguistik tertentu yang dapat dimanfaatkan untuk deteksi [23]. Penelitian lain juga menemukan bahwa teks buatan cenderung lebih formal dan kurang menggunakan gaya bahasa khas manusia seperti humor atau metafora [24].

RoBERTa merupakan hasil replikasi dan optimisasi ulang dari pretraining BERT yang dilakukan secara lebih sistematis, dengan skala data lebih besar, pemilihan hyperparameter yang lebih tepat, serta pelatihan tanpa penggunaan Next Sentence Prediction (NSP) [25]. RoBERTa bukan hanya mampu melampaui performa BERT, tetapi juga berhasil mencapai hasil *state-of-the-art* pada berbagai benchmark NLP ternama seperti GLUE, RACE, dan SQuAD [25]. Pendekatan berbasis model Transformer seperti RoBERTa telah digunakan secara luas dalam tugas atribusi kepenulisan, termasuk untuk membedakan teks buatan dari berbagai model generatif [9, 16]. Dalam salah satu studi [9], model RoBERTa-Sentinel mencapai 97% akurasi yang dievaluasi pada dua dataset utama: *OpenGPTText-Final* dan *OpenGPTText*. Pada dataset *OpenGPTText-Final*, model ini memperoleh nilai F1-score sebesar 0,94, dengan False Positive Rate (FPR) sebesar 9,0% dan False Negative Rate (FNR) sebesar 3,2%. Sementara pada dataset *OpenGPTText*, RoBERTa-Sentinel menghasilkan F1-score sebesar 0,89, FPR sebesar 21,6%, dan FNR sebesar 1,9%. Studi lain yang menggunakan model RoBERTa pada dataset berbeda juga menunjukkan performa tinggi, dengan precision sebesar 94,6%, recall 95,9%, F1-score 95,4%, dan akurasi 95,7% [26]. Model lain yang juga merupakan turunan dari RoBERTa, yaitu Distil-RoBERTa menunjukkan performa deteksi yang lebih baik dibandingkan pendekatan berbasis perplexity score [27].

Meskipun berbagai pendekatan telah dikembangkan, tantangan utama dalam deteksi teks buatan adalah meningkatnya kualitas keluaran model generatif, yang membuat perbedaan antara teks buatan dan teks manusia semakin tipis [16, 20]. Oleh karena itu, dibutuhkan pendekatan yang lebih adaptif dan berbasis pada model

yang kuat seperti RoBERTa, yang mampu menangkap nuansa linguistik secara lebih mendalam dan mempertahankan performa yang stabil meskipun dihadapkan pada model generatif canggih seperti ChatGPT [25, 28].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah disebutkan, berikut adalah rumusan masalah dari penelitian ini.

1. Bagaimana mengimplementasikan model *RoBERTa* untuk deteksi teks buatan *ChatGPT*?
2. Berapa tingkat *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dari model *RoBERTa* dalam mendeteksi teks buatan *ChatGPT*?

1.3 Batasan Permasalahan

Berikut merupakan batasan masalah dalam penelitian ini.

1. Data yang digunakan adalah dataset OpenGPTText yang berisi teks yang dihasilkan oleh model gpt-3.5-turbo (ChatGPT) dan dataset OpenWebText yang berisi teks yang ditulis oleh manusia.
2. Penelitian ini hanya berfokus pada deteksi teks berbahasa Inggris yang dihasilkan oleh ChatGPT, tidak mencakup bahasa lain atau model AI generatif lainnya.

1.4 Tujuan Penelitian

Berdasarkan rumusan dan batasan masalah yang telah diuraikan, tujuan dari penelitian ini adalah sebagai berikut.

1. Mengimplementasikan model *RoBERTa* yang dapat mendeteksi teks buatan *ChatGPT*.
2. Mengukur tingkat *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dari model *RoBERTa* dalam mendeteksi teks buatan *ChatGPT*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah sebagai berikut.

1. Memberikan pemahaman tentang efektivitas model RoBERTa dalam membedakan antara teks buatan dan teks manusia.
2. Menjadi dasar bagi pengembangan alat bantu atau aplikasi deteksi teks buatan AI di masa mendatang.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

1. Bab 1 PENDAHULUAN
   Bab ini membahas latar belakang masalah, merumuskan masalah yang diteliti, menetapkan batasan penelitian, menjelaskan tujuan yang ingin dicapai, serta menguraikan manfaat dari penelitian.
2. Bab 2 LANDASAN TEORI
   Bab ini menyajikan teori-teori yang mendasari penelitian, seperti Data Pre-processing, konsep Natural Language Processing, arsitektur Transformer, model RoBERTa, dan metrik evaluasi berupa *Confusion Matrix*.
3. Bab 3 METODOLOGI PENELITIAN
   Bab ini menjelaskan metodologi yang digunakan dalam penelitian, termasuk tahapan dan alur proses yang dilakukan untuk mencapai tujuan penelitian.
4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil pengujian dan analisis yang dilakukan terhadap model, serta mengevaluasi performa model berdasarkan data yang digunakan.
5. Bab 5 SIMPULAN DAN SARAN
   Bab ini merangkum temuan utama dari penelitian dan memberikan rekomendasi untuk penelitian selanjutnya yang dapat mengembangkan pendekatan atau model yang telah digunakan.