BAB 1
PENDAHULUAN

1.1 Latar Belakang

Dokumen seringkali digunakan para ahli dalam dua pengertian. Pertama, berarti sumber tertulis bagi informasi sejarah sebagai kebalikan dari kesaksian lisan, artefak, peninggalan-peninggalan terlukis, dan petilasan-petilasan arkeologis. Pengertian kedua diperuntukkan bagi surat-surat resmi dan surat-surat negara seperti surat perjanjian, undang-undang, hibah, konsesi, dan lainnya (Gottschalk, 1986). Tetapi dalam pembuatannya memerlukan ketelitian yang cukup baik agar dokumen yang dibuat tidak perlu dilakukan revisi. Merevisi dokumen merupakan pekerjaan yang memakan waktu karena pengerjaannya masih dilakukan secara manual oleh manusia.

Salah satu jenis dokumen yang cukup familiar di lingkungan akademis adalah dokumen jurnal. Menurut Lasa Hs (2009), "Jurnal ilmiah adalah publikasi ilmiah yang berisi tentang kegiatan terkait ilmu pengetahuan seperti pengamatan empiris, pengembangan gagasan, sampai kumpulan dari pengetahuan baru." Dan dalam perilisannya, jurnal ilmiah memiliki beberapa tahap yang cukup panjang dan memakan waktu, salah satu tahapannya adalah *"journal review" journal review* membutuhkan waktu yang cukup lama karena pengerjaannya masih dilakukan secara manual oleh pihak tim perilisan jurnal, sehingga dapat disimpulkan bahwa lamanya proses *journal review* disebabkan oleh banyaknya revisi pada sebuah jurnal.

Belum lama ini, fenomena ChatGPT sedang meningkat popularitasnya. Hal ini disebabkan oleh kemudahan akses dan kecanggihan yang dimiliki oleh ChatGPT, yang mampu membantu menyelesaikan banyak masalah. Teknologi dibalik ChatGPT adalah sebuah *language model* yang berukuran sangat besar untuk saat ini, yaitu *Generative Pre-Trained Transormers* (GPT) yang pertama kali dikenalkan pada 2018 oleh OpenAI pada sebuah artikel berjudul "*Improving Language Understanding by Generative Pre-Training*" GPT memiliki beberapa iterasi dan saat ini versi GPT sudah mencapai versi ke 4, GPT memiliki kemampuan dalam memahami konteks pada sebuah teks secara mendalam sehingga model tersebut dapat membantu menyelesaikan banyak permasalahan terkait dengan pemodelan bahasa.

Penelitian sebelumnya telah menunjukkan bahwa penggunaan teknologi NLP *(Natural Language Processing)* untuk memeriksa dan memperbaiki kesalahan penulisan sudah dilakukan. Sebagai contoh, penelitian yang dilakukan oleh (Musyafa, Gao, Solyman, Wu, & Khan, 2022). Dalam penelitiannya mengusulkan *Grammatical error correction* (GEC) otomatis berbasis arsitektur *Transformer* untuk memperbaiki kesalahan tata bahasa Indonesia dan juga dapat digunakan untuk memperbaiki teks *low resources languages* dan mencapai hasil yang signifikan mengungguli model GEC sebelumnya dengan skor F1 rata-rata 0,7194 dan skor BLEU 78,13, yang relatif efektif dalam tugas koreksi kesalahan tata bahasa.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dituliskan, pada penelitian ini mengangkat beberapa rumusan masalah, diantaranya:

- Apakah bisa menerapkan model GPT-2 untuk tugas koreksi kesalahan penulisan dokumen jurnal.
- Apa saja jenis kesalahan penulisan yang dapat dikoreksi menggunakan GPT-2.
- Seberapa efektif dan efisien penggunaan GPT-2 dalam mengoreksi kesalahan penulisan pada dokumen jurnal 6-8 halaman.

1.3 Tujuan Penelitian

Beberapa tujuan yang ingin dicapai setelah pengerjaan tugas akhir diantaranya:

- Menerapkan model GPT-2 untuk tugas koreksi penulisan dokumen jurnal.
- Mengidentifikasi terkait kesalahan penulisan apa saja yang dapat diperbaiki oleh GPT-2.
- Mengukur kinerja GPT-2 pada tugas koreksi penulisan dokumen jurnal.

1.4 Ruang Lingkup

Penelitian ini memiliki ruang lingkup sebagai berikut:

- Penerapan model akan dilakukan pada dokumen jurnal dengan 6-8 halaman dalam bahasa Indonesia.
- Bagian jurnal yang akan dikoreksi oleh model adalah hanya pada bagian teks.
- Teks yang akan dikoreksi dalam satu waktu oleh model tidak kurang dari 55 kata dan tidak lebih dari 512 token.
- Model akan menerima *input* dokumen jurnal dengan format PDF.
- *Dataset* yang dikumpulkan berasal dari jurnal yang ada di internet dengan tema atau topik pembahasan yang bebas (Tidak ditentukan).
- Varian jenis kesalahan penulisan yang ada pada *dataset* meliputi penggantian sinonim, pemberlakuan typo, pemberlakuan seluruh data menjadi *lowercase*, penerapan spasi setelah tanda baca yang tidak tepat, dan penggandaan kata secara acak pada setiap paragraf atau kalimat pada *dataset input*.

1.5 *State of The Art*

Tabel 1. 1 State of The Art

Judul Jurnal	**Pembahasan**
Attention Is All You Need **Peneliti:** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jakob Uszkoreit, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. **Tahun:** 2017	**Hasil Penelitian:** Artikel ini memperkenalkan arsitektur *Transformer* model transduksi urutan pertama yang sepenuhnya didasarkan pada perhatian (*attention*), menggantikan *layers reccurent* yang paling umum digunakan dalam arsitektur pengode-penyandi (*encoder-decoder*) dengan *multi-headed self-attention*. **Alasan Menjadi Tinjauan Penelitian:**

***Publisher*:** NIPS	Sebagai panduan dalam memahami arsitektur *Transformer*, sebab bagian arsitektur *decoder* digunakan pada GPT
Improving Language Understanding by Generative Pre-training **Peneliti:** Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever **Tahun:** 2018 ***Publisher*:** OpenAI	**Hasil Penelitian:** Artikel ini membahas tentang bagaimana pemahaman bahasa alami yang kuat melalui tahap *generative pre-training* dan *discriminative fine tuning*. Hasil menunjukkan peningkatan kinerja yang signifikan memang memungkinkan, dan memberikan petunjuk tentang model (*Transformer*) dan *dataset* (teks dengan ketergantungan jarak jauh) mana yang paling cocok dengan pendekatan ini. **Alasan Menjadi Tinjauan Penelitian:** Sebagai panduan dasar untuk pemahaman terkait *GENERATIVE PRE-TRAINED TRANSFORMERS*.
Language Models are Unsupervised Multitask Learners **Peneliti:** Alec Radford, JeffreyWu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever	**Hasil Penelitian:** Artikel ini menjelaskan ketika model bahasa besar dilatih pada *dataset* yang cukup besar dan beragam, model tersebut mampu berperforma dengan baik di banyak domain dan *dataset*. Keragaman tugas yang dapat dilakukan oleh model dalam pengaturan *zero-shot*

Tahun: 2019 ***Publisher*:** OpenAI	menunjukkan bahwa model berkapasitas tinggi yang dilatih untuk memaksimalkan peluang dari korpus teks yang cukup beragam mulai belajar bagaimana melakukan sejumlah tugas secara mengejutkan tanpa perlu supervisi eksplisit. **<u>Alasan Menjadi Tinjauan Penelitian:</u>** Penelitian ini membantu menjelaskan mengenai model yang mampu melakukan berbagai tugas secara *zero-shot*, yaitu tanpa pelatihan khusus pada tugas tertentu. Hal ini menunjukkan fleksibilitas dan adaptabilitas model dalam memanfaatkan pengetahuannya untuk berbagai konteks dan tugas.
The survey: Text generation models in deep learning **Peneliti:** Touseef Iqbal, Shaima Qureshi **Tahun:** 2020	**<u>Hasil Penelitian:</u>** Jurnal ini membahas secara singkat perkembangan yang terjadi dalam pemodelan *Deep Generative*. Jurnal ini menyimpulkan bahwa di dalam bidang data kontinu (gambar), dominasi dipegang oleh GAN sementara di dalam bidang data diskrit (teks), dominasi dipegang oleh Variational Auto-*Encoder*s.

***Publisher*:** King Saud University	**<u>Alasan Menjadi Tinjauan Penelitian:</u>** Membantu dalam memahami pengertian mengenai model *generative* teks.
Automatic Correction of Indonesian Grammatical Errors Based on Transformers **Penulis:** Ahmad Musyafa, Ying Gao, Aiman Solyman, Chaojie Wu, Siraj Khan **Tahun:** 2022 ***Publisher*:** MDPI	**<u>Hasil Penelitian:</u>** Artikel ini membahas model GEC berbasis *neural-based* pertama untuk bahasa Indonesia yang juga bisa memperbaiki teks bahasa-bahasa dengan sumber daya terbatas lainnya. Selain itu, Untuk mengatasi masalah kekurangan data dalam GEC bahasa Indonesia, artikel ini mengusulkan metode *confusion* semi *supervised* untuk menghasilkan data pelatihan paralel dari korpus monolingual di luar domain. Hasil eksperimen menunjukkan keunggulan dengan skor F1 0.7194 dan BLEU 78.13, efektif dalam koreksi kesalahan tata bahasa, tetapi masih belum menangani kesalahan semantik dan sintetik. **<u>Alasan Menjadi Tinjauan Penelitian:</u>** Artikel ini memberi pengetahuan mengenai tugas koreksi kesalahan tata bahasa Indonesia.

1.6 Sistematika Penulisan

Secara garis besar penulisan laporan tugas akhir ini terbagi dalam beberapa bab yang tersusun antara lain sebagai berikut:

1. BAB I PENDAHULUAN

   Bab ini berisi latar belakang, rumusan masalah, tujuan penelitian, ruang lingkup, *state of the art*, dan sistematika penulisan.

2. BAB II LANDASAN TEORI

   Pada bab ini diuraikan tentang teori yang berhubungan dengan menunjang penulisan tugas akhir.

3. BAB III METODOLOGI PENELITIAN

   Pada bab berisi penjelasan tentang metodologi penelitian yang digunakan, permasalahan yang ada serta pengerjaan yang dilakukan. Implementasi yang dilakukan menggunakan model GPT-2.

4. BAB IV HASIL DAN PEMBAHASAN

   Pada bab ini dijelaskan tentang *software/hardware environment* dan hasil implementasi model GPT-2 *small* untuk tugas koreksi dokumen jurnal.

5. BAB V PENUTUP

   Pada bab ini berisi tentang kesimpulan yang diperoleh dari hasil pengujian model, serta saran perbaikan dan pengembangan model ke depannya.