BAB 1 PENDAHULUAN

1.1. Latar Belakang

Bahasa merupakan salah satu unsur penting yang tidak dapat terpisahkan dari kehidupan sehari-hari manusia. Dalam segala aktivitasnya, manusia senantiasa berinteraksi menggunakan bahasa. Peran bahasa sangat penting dalam kehidupan manusia sebagai sarana komunikasi untuk menyampaikan berbagai informasi [1], [2]. Setiap bangsa memiliki ciri khas yang berbeda-beda, termasuk dalam bahasa yang digunakan untuk berkomunikasi dengan beragam variasi tergantung pada konteksnya. Tidak ada dua bahasa di dunia ini yang serupa. Setiap manusia di bumi ini memiliki bahasa masing-masing yang digunakan untuk berkomunikasi. Oleh karena itu, tidaklah tepat jika seseorang menyalahkan atau memaksa setiap individu untuk selalu menggunakan Bahasa Indonesia [3].

Bahasa memiliki dua bentuk, yaitu lisan dan teks. Lisan dapat dibentuk dengan cara mengucapkan, sementara teks dengan menulis. Menulis merupakan keterampilan pembelajaran yang menantang bagi individu meskipun menggunakan bahasa ibu sendiri, yakni Bahasa Indonesia. Supaya dapat dipahami oleh sesama manusia, penulisan teks bahasa harus memiliki struktur kalimat yang baik dan benar karena ia merupakan salah satu unsur penting dalam bahasa. Setiap bahasa memiliki struktur kalimatnya masing-masing dan setiap kalimat terdiri dari satu atau lebih klausa [4]. Kesalahan dalam menulis dapat terjadi seperti halnya dalam tanda baca, ejaan, dan pemilihan kata. Namun seorang penulis juga sering menghadapi kesulitan tambahan dalam menyusun teks yang memiliki tata bahasa yang jelas dan mudah dipahami [1], [2]. Kesalahan ini dikenal dengan kesalahan sintaksis. Sintaksis adalah cabang ilmu bahasa yang mempelajari penulisan dalam kalimat. Kesalahan dalam bidang sintaksis biasanya terjadi pada tingkat frasa dan kalimat. Pada tingkat frasa, kesalahan dapat meliputi penggunaan

preposisi yang tidak tepat, susunan kata yang tidak benar, penggunaan unsur yang berlebihan atau mubazir, penjamakan ganda, penggunaan bentuk superlatif yang berlebihan, serta kesalahan dalam penggunaan bentuk resiprokal. Sedangkan pada tingkat kalimat, kesalahan dapat berupa kalimat tanpa subjek, penyisipan antara subjek dan predikat, kalimat yang tidak logis, penggunaan konjungsi yang berlebihan, penggunaan kata tanya yang tidak perlu, serta penggunaan istilah asing [5], [6].

Salah satu prinsip dalam gaya penulisan, khususnya di bidang akademik adalah konsistensi. Konsistensi dalam penulisan mencakup penggunaan tata bahasa yang seragam, seperti penggunaan t*enses* yang benar, kata baku, serta diksi yang konsisten dengan mengikuti pedoman aturan tata bahasa yang telah ditetapkan oleh Tata Bahasa Baku Indonesia (TBBI). Hal ini merupakan salah satu bentuk memahami dasar-dasar Bahasa Indonesia [7], [8]. Dengan tidak mengikuti pedoman TBBI, permasalahan dalam penulisan kalimat sering kali menimbulkan kebingungan dan kesalahpahaman di antara para pembaca saat membaca teks berbahasa Indonesia [9].

Dalam kehidupan sehari-hari, kasus kesalahan penulisan tata bahasa Indonesia, khususnya dalam bidang sintaksis, masih sering muncul di media, baik elektronik maupun cetak, seperti surat kabar harian atau media berita daring (*online*). Hal ini dikarenakan dalam penulisannya, penulis masih dapat menyebabkan berbagai kesalahan (*human error*), seperti kesalahan pengetikan, kurangnya ketelitian dalam proses penyuntingan, dan penggunaan kaidah kebahasaan yang tidak tepat. Kesalahan penulisan kalimat baku bahasa Indonesia yang terjadi juga dapat diakibatkan oleh penulis yang belum memahami standar penulisan Bahasa Indonesia yang baik dan benar sesuai dengan TBBI. Kesalahan tersebut tentu saja mengurangi kualitas teks dan maknanya. Akibatnya, kesalahan pemahaman pembaca dalam memahami penulisan tersebut dapat terjadi [10].

Fenomena-fenomena tentang kesalahan tata bahasa dalam penulisan teks berbahasa Indonesia juga sering terjadi di beberapa kawasan sekolahan

dan perguruan tinggi Indonesia. Hal ini ditemukan dalam hasil analisis dari beberapa penelitian. Terdapat sebuah penelitian yang menganalisis kesalahan sintaksis tata bahasa di *caption* konten Instagram Kementerian Kesehatan Republik Indonesia (Kemenkes). Penelitian tersebut menunjukkan hasil dari analisis 40 *caption* konten Ins*tag*ram dari akun resmi Kemenkes. Hasil tersebut menunjukkan 40 kesalahan sintaksis yang terdiri dari 14 kesalahan frasa dan 26 kesalahan kalimat. Kesalahan frasa terjadi akibat penggunaan preposisi yang keliru, susunan kata yang tidak pas, dan unsur yang mubazir. Sementara itu, kesalahan kalimat mencakup ketiadaan subjek, kalimat yang tidak tuntas, adanya kata yang menyisip di antara predikat dan objek, penggunaan konjungsi yang hilang atau berlebihan, serta pemakaian istilah asing dan kata tanya yang tidak perlu [11].

Penelitian selanjutnya juga membahas tentang analisis kesalahan sintaksis bahasa Indonesia dalam teks eksposisi dari siswa kelas X di SMK Negeri 8 Palembang. Hasil penelitian tersebut menunjukkan bahwa terdapat kesalahan penggunaan sintaksis berupa frasa sebanyak 30 kalimat (32,98%) dan berupa kalimat sebanyak 63 kalimat (67,02%) dari 30 teks eksposisi tugas bahasa Indonesia yang dikerjakan oleh para siswa. Kesalahan penggunaan frasa tersebut meliputi penggunaan preposisi yang tidak tepat, susunan kata yang tidak tepat, penggunaan unsur yang berlebihan atau mubazir, penggunaan bentuk superlatif yang berlebihan, penjamakan yang ganda, dan penggunaan bentuk resiprokal yang salah. Sementara itu, kesalahan penggunaan struktur kalimat tersebut meliputi kalimat yang tidak berpredikat, kalimat buntung, kalimat yang tidak logis, penggunaan kata tanya yang tidak perlu, urutan yang tidak paralel, penghilangan konjungsi, dan penggunaan konjungsi yang berlebihan [12].

Penelitian selanjutnya membahas tentang analisis kesalahan morfologi pada narasi berita daring dari kompasiana.com edisi Februari 2023. Hasil penelitian tersebut menunjukkan bahwa dari 15 sampel artikel

kompasiana.com edisi Februari 2023, ditemukan 26 kasus (70,27%) penggunaan afiks yang keliru dan 11 kasus (29,73%) penghilangan afiks [13].

Penelitian selanjutnya membahas tentang analisis kesalahan morfologi dan sintaksis pada teks ulasan karya siswa MTs Negeri 5 Ponorogo tahun 2020/2021. Dari 15 teks ulasan, ditemukan 86 data yang memiliki kesalahan. Kesalahan-kesalahan tersebut mencakup 51 data kesalahan morfologi dan 35 data dari kesalahan sintaksis [14].

Selain itu, terdapat juga penelitian yang serupa dengan studi kasus skripsi mahasiswa di Universitas HKBP Nomensen Pematang Siantar, Fakultas Keguruan dan Ilmu Pendidikan, Program Studi Pendidikan Bahasa Indonesia. Hasil penelitian tersebut menunjukkan bahwa dari 40 dokumen skripsi mahasiswa angkatan 2018 yang lulus tahun 2022, terdapat kesalahan struktur frasa yang berupa penggunaan unsur yang berlebihan (mubazir) sebanyak 150 data dan ketidaktepatan dalam menggunakan preposisi sebanyak 134 data. Selain itu, terdapat kesalahan unsur kalimat yang didominasi oleh kalimat tanpa subjek sebanyak 528 data, penggunaan istilah asing sebanyak 234 data, dan penggunaan kata tanya yang tidak perlu sebanyak 75 data [15].

Secara keseluruhan, kelima penelitian yang dirangkum di atas menganalisis bahwa lebih dari 140 dokumen berbeda, mulai dari unggahan media sosial, artikel daring, hingga karya tulis siswa dan skripsi. Dari analisis gabungan ini, teridentifikasi total 1.377 kesalahan tata bahasa. Data ini menunjukkan dominasi kesalahan sintaksis yang mencapai 1.289 kasus (sekitar 93,6%), sementara kesalahan morfologi tercatat sebanyak 88 kasus (sekitar 6,4%). Angka-angka ini secara kuantitatif mengonfirmasi bahwa persoalan struktur kalimat dan frasa menjadi tantangan utama dalam penulisan Bahasa Indonesia di berbagai tingkatan.

Fenomena-fenomena yang terjadi di atas menunjukkan bahwa persoalan kesalahan tata bahasa dalam penulisan teks bahasa Indonesia

yang terjadi di berbagai jenjang pendidikan dan media daring bersifat sistemik dan meluas, baik dari aspek morfologi maupun sintaksis. Hal tersebut dibuktikan dengan adanya ratusan kesalahan dalam struktur frasa, kalimat, dan penggunaan afiks dalam teks karya siswa, media sosial resmi, dan karya ilmiah mahasiswa. Kesalahan-kesalahan tersebut meliputi ketiadaan subjek atau predikat, struktur kalimat yang tidak logis, penggunaan preposisi atau konjungsi yang tidak tepat, redudansi unsur kalimat, hingga kehilangan atau kelebihan konjungsi. Fakta-fakta ini menunjukkan keurgenan penelitian ini untuk mengembangkan solusi yang sistematis dan adaptif dalam mengidentifikasi kesalahan tata bahasa secara otomatis.

Untuk mengatasi masalah tersebut, manusia sudah menerapkan sistem koreksi kesalahan tata bahasa tradisional dengan cara mendeteksi dan mengoreksinya secara manual, akan tetapi sering terjadi *human error* yang mengakibatkan akurasi yang didapatkan rendah dan waktu yang dibutuhkan cukup lama meskipun Bahasa Indonesia merupakan bahasa yang relatif tidak kompleks, baik bagi penutur asli maupun asing. Bahasa Indonesia memiliki struktur yang cukup sederhana dibandingkan dengan beberapa bahasa lain. Tidak ada konjugasi verba yang kompleks atau perubahan bentuk kata berdasarkan kasus atau gender. Dalam Bahasa Indonesia, verba tetap sama tanpa perubahan bentuk untuk menunjukkan orang, jumlah, atau waktu [8].

Saat ini, terdapat alat komersial yang efektif dalam mendeteksi dan memperbaiki kesalahan ejaan. Namun, alat untuk mendeteksi kesalahan tata bahasa dalam tulisan bahasa Indonesia, masih menjadi tantangan yang rumit dan belum ada solusi yang sepenuhnya matang. Jika komputer dapat dengan cepat mengenali kesalahan tata bahasa secara mendalam, maka pengalaman belajar Bahasa Indonesia akan menjadi lebih baik. Contohnya adalah aplikasi SIPEBI. SIPEBI merupakan platform asistensi penulisan berbasis bahasa Indonesia akan tetapi, SIPEBI hanya menargetkan peningkatan kosa kata dan ejaan. SIPEBI tidak menargetkan terhadap perbaikan struktur kalimat. Dengan kata lain SIPEBI tidak memiliki

kemampuan untuk mengenali kesalahan dalam konstruksi bahasa dan cacat kalimat. Dalam peningkatan kosa kata dan ejaan pun, SIPEBI masih belum dapat melakukan penyuntingan dengan rinci dan khusus, seperti pada perbaikan huruf kapital. SIPEBI dapat mengidentifikasi kesalahan penggunaan huruf kapital pada awal kalimat, tetapi tidak dapat mendeteksi kesalahan serupa pada penggunaan huruf kapital dalam nama orang, tempat, dan sebagainya [16].

Contoh lain dari platform asistensi penulisan lain adalah Grammarly. Grammarly sudah menggunakan pendekatan *machine learning* tingkat lanjut untuk membuka jalan baru dalam pemrosesan bahasa alami yang menganalisis kalimat tertulis untuk memahami konteks dan nada. Dalam konteks tata bahasa, Gramarly memiliki tiga fitur yang dapat digunakan sebagai asistensi penulisan, seperti *grammar checker*, *plagiarism checker*, dan *essay checker*. Sayangnya, platform ini hanya mendukung bahasa Inggris [17].

Masalah-masalah di atas menjadi titik berat pada penelitian ini. Salah satu penyelesaian masalah-masalah tersebut adalah mengembangkan teknologi pemrosesan bahasa alami (*Natural Language Processing*) berbasis model *deep learning* sesuai dengan konteks penelitian ini. Pengembangan model tersebut dilakukan dengan menggunakan arsitektur Transformer, yakni dengan menerapkan metode pengecekan tata bahasa berbasis aturan yang sesuai. Maka dari itu, pendeteksian kesalahan dalam kalimat Bahasa Indonesia dapat dilakukan lebih mendalam [18].

Jaringan saraf tiruan *multilayer* dari Transformer digunakan untuk membangun model bahasa guna menilai apakah kata-kata yang terbentuk menjadi sebuah kalimat merupakan kalimat normal, sehingga dapat mendeteksi kesalahan dalam tata bahasa kalimat tersebut [18]. Penggunaan model Transformer dalam pengecekan tata bahasa menjadi solusi yang relevan. Hal ini memungkinkan model untuk belajar dari pola-pola tata bahasa dalam teks, memahami konteks kalimat, serta menangani variasi

dalam penggunaan kata dan struktur kalimat yang beragam. Dengan demikian, penggunaan model Transformer dalam mendeteksi kesalahan tata bahasa pada teks berbahasa Indonesia diharapkan dapat membantu meningkatkan kualitas teks secara efisien dan efektif.

Transformer telah diadopsi secara luas di berbagai bidang, seperti *Natural Language Processing* (NLP), *Computer Vision* (CV), dan pemrosesan suara. Dalam bidang NLP, Transformer awalnya diusulkan sebagai model *sequence-to-sequence* untuk *machine translation* (terjemahan mesin). Hal ini terbukti bahwa Transformer dapat dilatih dengan lebih cepat daripada arsitektur berbasis rekuren atau konvolusi, seperti RNN dan CNN, dan telah mencapai hasil yang sangat baik dalam tugas *machine translation*, seperti pada tugas terjemahan Inggris-Jerman dan Inggris-Perancis [19]. Setelah itu, Transformer mulai berkembang menjadi *pre-trained* berbasis Transformer (PTM) yang dapat melakukan berbagai macam tugas. Pada penelitian selanjutnya yang dilakukan pada 2020, Qiu dkk. telah menunjukkan survei bahwa model *pre-trained* berbasis Transformer (PTM) dapat mencapai performa terbaik dalam berbagai tugas. Akibatnya, Transformer telah menjadi arsitektur utama dalam NLP, terutama untuk PTM [20], [21].

Dalam kasus pendeteksian kesalahan tata bahasa Indonesia maupun bahasa asing menggunakan arsitektur jaringan saraf tiruan, ada beberapa peneliti yang telah melakukan penelitian tersebut. Hal ini melibatkan proses klasifikasi teks terlebih dahulu untuk membandingkan antara kalimat yang benar dan salah. Contoh arsitektur yang pernah digunakan untuk klasifikasi tersebut adalah seperti, dan RNN [22]. Dibandingkan dengan arsitektur tersebut, peneliti ingin mengembangkannya dengan arsitektur berbasis *self-attention*, yaitu Transformer yang terkenal mutakhir dalam NLP, apalagi setelah kemunculan *pre-trained* model BERT (*Bidirectional* Encoder Representations from Transformers. Pada 2018, Tim Peneliti Google, yakni Devlin dkk. memperkenalkan BERT yang dirancang untuk melatih representasi *bidirectional* yang mendalam dari teks tanpa label dengan

secara bersamaan mengondisikan konteks kiri dan kanan di semua lapisan. Akibatnya, model BERT yang telah dilatih sebelumnya dapat disesuaikan hanya dengan satu lapisan *output* tambahan untuk menciptakan model mutakhir untuk berbagai tugas (*downstream tasks*), seperti tanya jawab (*Qusetion Answering*) dan inferensi bahasa (*Natural Language Inference*), tanpa perlu modifikasi arsitektur spesifik untuk setiap tugas. Akhirnya, BERT dapat memahami makna kata dan kalimat (*Natural Language Understanding*) dalam konteks yang lebih luas, sehingga menghasilkan hasil yang lebih akurat dan informatif. BERT berhasil mencapai *state-of-the-art* (hasil terbaik) sampai saat ini [23].

Selain itu, ada juga platform yang telah sukses dalam bidang NLP yang memanfaatkan Transformer sebagai dasarnya, yakni ChatGPT (Chat Generative *Pre-trained* Transformer). Dengan menggunakan model GPT sebagai dasarnya, ChatGPT memiliki potensi untuk mengubah cara manusia berinteraksi dengan komputer dan mesin dengan komunikasi yang lebih alami dan intuitif. Akibat dari hal tersebut, ChatGPT mewakili terobosan yang signifikan dalam bidang NLP. ChatGPT telah merevolusi NLP dengan menghasilkan teks mirip manusia, lengkap dengan konteks dan koherensi [24].

Berdasarkan latar belakang yang tertulis di atas, peneliti ingin melakukan penelitian tentang pengembangan model berbasis arsitektur Transformer untuk mendeteksi kesalahan tata bahasa Indonesia.

1.2. Rumusan Masalah

- Bagaimana mengimplementasikan arsitektur Transformer *Encoder-Only* pada model untuk mendeteksi kesalahan tata bahasa Indonesia?
- Berapa tingkat akurasi model dalam mengukur kinerja dan evaluasi model dalam mendeteksi kesalahan tata bahasa Indonesia?
- Bagaimana hasil perbandingan arsitektur Transformer dengan jumlah *encoder layer* yang berbeda?

1.3. Tujuan dan Manfaat

Berdasarkan rumusan masalah yang telah diuraikan, penelitian ini memiliki tiga tujuan utama. Pertama, penelitian ini bertujuan untuk mengimplementasikan sebuah model yang didasarkan pada arsitektur Transformer. Implementasi model ini merupakan realisasi awal dari upaya pencegahan dan deteksi kesalahan tata bahasa dalam teks berbahasa Indonesia. Kedua, penelitian ini bertujuan untuk mengukur tingkat akurasi model yang telah diimplementasikan tersebut. Pengukuran ini difokuskan pada evaluasi kinerja model dalam mendeteksi kesalahan tata bahasa Indonesia. Ketiga, penelitian ini bertujuan untuk membandingkan antara hasil dari evaluasi model Transformer dengan jumlah *encoder layer* yang berbeda. Adapun manfaat dari penelitian ini yakni sebagai berikut:

a. Bagi Peneliti

Manfaat yang bisa didapatkan oleh peneliti adalah dapat memperluas pemahaman ilmiah, khususnya dalam implementasi arsitektur Transformer untuk *Grammatical Error Detection* (GED) dalam bahasa Indonesia. Selain itu, penelitian ini juga dapat dijadikan referensi penelitian selanjutnya berdasarkan model yang telah dikembangkan.

b. Bagi Penulis

Adapun manfaat yang bisa didapatkan oleh para penulis adalah dapat meningkatkan kualitas teks secara efisien dan efektif. Dengan teknologi ini, sistem dapat belajar dari pola-pola tata bahasa dalam teks, memahami konteks kalimat, dan menangani jenis-jenis kesalahan dalam struktur kalimat yang beragam. Hal ini memungkinkan deteksi kesalahan tata bahasa yang lebih mendalam dan akurat, sehingga dapat membantu penulis untuk menghasilkan teks yang lebih koheren dan mudah dipahami oleh para pembacanya.

1.4. Batasan Masalah

Dalam penelitian ini, terdapat beberapa batasan masalah yang akan diterapkan. Berikut adalah batasan-batasan masalah yang akan diterapkan pada penelitian ini:

- Penelitian ini hanya berfokus terhadap perancangan model berbasis arsitektur Transformer dengan hanya menggunakan *encoder layer* dalam mendeteksi kesalahan tata bahasa Indonesia.
- Tata bahasa yang diteliti terbatas pada teks tertulis dalam kalimat bahasa Indonesia.
- Deteksi kesalahan hanya sebatas tingkat kalimat secara sintaksis dan morfologis. Kesalahan tersebut terdiri dari tujuh jenis, yakni kesalahan susunan kata, formasi jamak, penggunaan imbuhan (afiks), preposisi, konjungsi, subjek yang hilang, dan predikat yang hilang. Hal ini tidak termasuk dalam tingkat ejaan.
- Aturan tata bahasa merujuk pada aturan penulisan kalimat secara sintaksis dan morfologis dalam TBBI (Tata Bahasa Baku Indonesia).
- Evaluasi model berfokus pada nilai akurasi dan *F1-score*.

1.5. Metode Penelitian

Penelitian ini dilakukan dengan pendekatan studi literatur untuk mengumpulkan teori terkait arsitektur Transformer dan aturan tata bahasa Indonesia, pengumpulan data, *pre-procesing data*, perancangan model, dan implementasi model yang dimulai dari pelatihan (*training*), validasi (*validation*), dan pengujian (*testing*).