BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi yang pesat telah mendorong transformasi signifikan dalam lanskap jurnalisme, khususnya dengan berkembangnya media berita daring yang menuntut kecepatan dan aksesibilitas penyampaian informasi [1]. Namun, di balik kecepatan tersebut, wartawan sering kali dihadapkan pada beban kerja yang tinggi (*over capacity*) yang memicu kelelahan fisik maupun mental sehingga memengaruhi performa kerja mereka [2]. Situasi ini menempatkan jurnalis dalam dilema antara memprioritaskan kecepatan tayang atau menjaga akurasi berita [3]. Akibatnya, tekanan tenggat waktu yang ketat sering kali menyebabkan terjadinya kelalaian dalam proses penulisan [4]. Kondisi penulisan yang terburu-buru demi memenuhi batasan waktu ini menyisakan sedikit ruang untuk evaluasi, sehingga memunculkan berbagai kesalahan linguistik dan struktur kebahasaan pada berita yang dihasilkan [5]. Kondisi tersebut menunjukkan bahwa proses pemeriksaan bahasa secara manual semakin sulit dilakukan secara konsisten, terutama ketika jumlah artikel yang harus diproduksi dan ditinjau terus meningkat. Oleh karena itu, diperlukan dukungan teknologi yang mampu membantu proses analisis dan pengendalian kualitas bahasa secara otomatis.

Dalam ekosistem digital ini, teknologi *Artificial Intelligence* (AI) dan *Natural Language Processing* (NLP) semakin banyak diadopsi untuk mendukung berbagai aspek komunikasi digital, mulai dari sistem temu kembali informasi, mesin penerjemah, hingga perangkat koreksi otomatis [6]. Di tengah tuntutan kecepatan publikasi, ketepatan struktur kalimat dalam teks berita menjadi aspek krusial yang menentukan kejelasan informasi, keterbacaan, dan kredibilitas media [7]. Kesalahan dalam penggunaan bahasa, sekecil apa pun, dapat menyebabkan pergeseran makna, ambiguitas informasi, dan pada akhirnya mencederai prinsip-prinsip *good journalism* [3][8]. Oleh karena itu, penerapan teknologi NLP untuk memastikan kualitas bahasa dalam berita daring menjadi kebutuhan strategis guna menjaga standar informasi yang dikonsumsi masyarakat luas.

Bahasa Indonesia memiliki karakteristik sintaksis yang khas, di mana struktur kalimat tersusun atas satuan-satuan bahasa berupa kata, frasa, dan

klausa yang memiliki peran dan fungsi spesifik seperti subjek, predikat, objek, pelengkap, dan keterangan [9][10]. Meskipun fleksibilitas struktur kalimat bahasa Indonesia memberikan keleluasaan ekspresi, kondisi ini justru menjadi tantangan tersendiri dalam konteks penulisan berita daring yang dikerjakan di bawah tekanan waktu dan volume produksi yang tinggi [11]. Berbagai kesalahan struktur kalimat atau sintaksis kerap ditemukan pada teks berita, antara lain kesalahan penempatan keterangan, ketidaksesuaian subjek-predikat, ambiguitas frasa, paralelisme yang tidak konsisten, serta kesalahan penggunaan pemarkah klausa majemuk [8][11]. Kesalahan-kesalahan tersebut dapat menyebabkan hilangnya sebagian hingga seluruh makna kalimat, yang berpotensi menimbulkan miskomunikasi dan menurunkan kepercayaan pembaca terhadap media [8]. Volume produksi berita yang sangat besar, dengan ratusan hingga ribuan artikel yang dapat diunggah setiap hari oleh satu portal berita, menyebabkan proses analisis dan penyuntingan manual menjadi kurang efisien, rentan terhadap ketidakkonsistenan, serta sulit diskalakan, sehingga beban kerja jurnalis dan editor dalam menjaga kualitas pemberitaan semakin meningkat [11].

Keberadaan kesalahan sintaksis pada media berita daring juga telah dibuktikan dalam berbagai penelitian. Analisis terhadap berita CNN Indonesia menunjukkan masih ditemukannya kesalahan sintaksis dalam penyusunan kalimat yang dipublikasikan secara daring [12]. Selain itu, penelitian pada teks berita Tribun News menunjukkan bahwa sebagian kalimat masih belum memenuhi prinsip kalimat efektif dan mengandung permasalahan pada struktur kalimat yang berpotensi mengurangi kejelasan informasi yang diterima pembaca [13]. Temuan tersebut menunjukkan bahwa kualitas struktur kalimat dalam berita daring masih perlu ditingkatkan agar informasi dapat disampaikan secara lebih jelas dan mudah dipahami. Oleh karena itu, diperlukan suatu teknologi yang mampu melakukan deteksi dan koreksi kesalahan sintaksis secara otomatis dalam proses penyuntingan berita. Pengembangan teknologi tersebut didasarkan pada kebutuhan untuk menjaga kualitas dan kredibilitas informasi di tengah peningkatan volume serta kecepatan publikasi berita daring, sehingga pemeriksaan struktur kalimat secara manual menjadi kurang efektif untuk dilakukan secara konsisten.

Untuk mengatasi permasalahan tersebut, diperlukan pendekatan otomatis yang mampu mendeteksi sekaligus mengoreksi kesalahan struktur kalimat pada teks berita. Integrasi metode berbasis aturan dan *deep learning* dinilai mampu memanfaatkan keunggulan kedua metode dalam menganalisis aspek sintaksis dan konteks kalimat [14]. *Context Free Grammar* (CFG) merupakan metode

berbasis aturan formal yang efektif dalam merepresentasikan struktur sintaksis kalimat secara hierarkis guna mengidentifikasi cacat sintaksis melalui proses parsing. Namun, pendekatan berbasis aturan sering kali kaku dan hanya mampu mengidentifikasi ada-tidaknya pelanggaran struktur tanpa dapat menghasilkan kalimat perbaikan secara otomatis [15]. Oleh karena itu, pemanfaatan model deep learning berbasis Transformer, khususnya *Text-to-Text Transfer Transformer* (T5), berperan sebagai pelengkap terhadap pendekatan berbasis aturan (*rule-based*). Dalam penelitian ini, metode *Context Free Grammar* (CFG) dan Stanza NLP difokuskan untuk tugas pelabelan kelas kata dan deteksi struktur, sedangkan model *Text-to-Text Transfer Transformer* (T5) diterapkan untuk melakukan koreksi otomatis terhadap struktur kalimat, meliputi perbaikan susunan unsur kalimat, penggunaan konjungsi, dan tanda baca berdasarkan konteks semantik, dengan tetap mempertahankan makna asli kalimat [16].

Sejak tahun 2020, platform U-Tapis telah mengembangkan berbagai modul deteksi kesalahan kebahasaan bahasa Indonesia yang mencapai akurasi tinggi, seperti modul kata depan di (LSTM, 96,46%), kata terikat (Rabin-Karp + Random Forest, 86,24%), peluluhan kata (Damerau-Levenshtein + BERT, 96,79%), dan kata baku (Levenshtein, 99%), dengan tiga modul utama telah memperoleh paten pada tahun 2025 [17][18][19]. Meskipun demikian, studi terdahulu terkait modul sintaksis kalimat pada tahun 2023 yang menggunakan Conditional Random Field (CRF) dan CFG masih menunjukkan keterbatasan, dengan akurasi 91% pada dataset terbatas (1.341 kalimat) [20]. Model CRF yang digunakan bergantung pada *feature engineering* manual dan asumsi Markov yang membatasi kemampuan menangkap dependensi kontekstual jangka panjang, serta hanya menghasilkan output biner tanpa memberikan identifikasi kategori kesalahan maupun usulan perbaikan kalimat secara otomatis. Akibatnya, proses koreksi masih memerlukan intervensi manual dari pengguna. Keterbatasan tersebut membuka peluang pengembangan model sintaksis generasi kedua yang tidak hanya mampu mendeteksi kesalahan secara lebih rinci, tetapi juga menghasilkan koreksi kalimat secara otomatis melalui pemanfaatan arsitektur Transformer berbasis T5.

Berdasarkan uraian tersebut, penelitian ini difokuskan pada pengembangan modul deteksi kesalahan struktur kalimat berbasis *Context Free Grammar* (CFG) serta modul koreksi kalimat berbasis *Text-to-Text Transfer Transformer* (T5) pada berita daring berbahasa Indonesia dengan memanfaatkan korpus 10.000 artikel Tribunnews yang dianotasi secara sintaksis. Penelitian ini merupakan bagian dari pengembangan modul sintaksis U-Tapis generasi kedua yang bertujuan mengatasi

keterbatasan pendekatan *Conditional Random Field* (CRF) pada penelitian sebelumnya. Model yang dikembangkan diharapkan mampu mendeteksi delapan kategori kesalahan struktur kalimat, seperti salah posisi keterangan, ketidaksesuaian subjek-predikat, dan kesalahan penggunaan pemarkah klausa majemuk, serta menghasilkan koreksi kalimat secara otomatis.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi dan koreksi kesalahan sintaksis bahasa Indonesia secara otomatis. Selain itu, hasil penelitian ini diharapkan dapat mendukung pengembangan modul sintaksis pada platform U-Tapis dan menjadi referensi bagi penelitian selanjutnya pada bidang *Natural Language Processing* untuk bahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka perumusan masalah dalam penelitian ini adalah:

1. Bagaimana penerapan Context Free Grammar (CFG) dan Text-to-Text Transfer Transformer (T5) dalam mendeteksi serta mengoreksi kesalahan struktur kalimat pada teks berita daring berbahasa Indonesia?
2. Bagaimana kemampuan modul CFG dalam mengidentifikasi delapan kategori kesalahan struktur kalimat serta kemampuan modul IndoT5 dalam menghasilkan koreksi kalimat yang sesuai dengan kaidah bahasa Indonesia?
3. Bagaimana kinerja modul CFG berdasarkan metrik accuracy, precision, recall, dan F1-score serta kinerja modul IndoT5 berdasarkan metrik Exact Match, ROUGE-L, dan GLEU?

1.3 Batasan Permasalahan

Agar penelitian ini lebih terarah dan fokus pada target yang ingin dicapai, penulis menetapkan batasan-batasan masalah sebagai berikut:

1. Sumber Data: Penelitian ini menggunakan korpus berisi 10.000 artikel berita daring berbahasa Indonesia yang dikumpulkan dari Tribunnews.
2. Lingkup Deteksi: Penelitian ini secara khusus dibatasi pada pengujian validitas dan analisis struktur tata kalimat. Pendeteksian kesalahan pada level

karakter, seperti ejaan (*typo*) dan imbuhan (*morfologi*), sepenuhnya berada di luar batasan sistem.

3. Kategori Kesalahan: Kategori kesalahan sintaksis yang dianalisis dalam penelitian ini terdiri atas delapan kategori, yaitu posisi keterangan menyela (K01), preposisi awal subjek (K02), hilang predikat (K03), konjungsi ganda (K04), konjungsi intrakalimat di awal kalimat (K05), comma splice (K06), keterangan aposisi (K07), dan subjek sama ganda (K08).

4. Sistem dalam penelitian ini terdiri atas dua modul utama, yaitu modul deteksi kesalahan struktur kalimat berbasis *Context Free Grammar* (CFG) dan modul koreksi kalimat berbasis *Text-to-Text Transfer Transformer* (T5). Kedua modul tersebut dirancang dan dievaluasi khusus untuk mendeteksi serta mengoreksi kesalahan struktur kalimat pada teks berita daring berbahasa Indonesia, dengan cakupan delapan kategori kesalahan sintaksis (K01–K08).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menghasilkan rancangan dan implementasi modul deteksi kesalahan struktur kalimat berbasis Context Free Grammar (CFG) serta modul koreksi kalimat berbasis Text-to-Text Transfer Transformer (T5).

2. Menganalisis kemampuan modul CFG dalam mengidentifikasi delapan kategori kesalahan struktur kalimat serta kemampuan modul IndoT5 dalam menghasilkan koreksi kalimat yang sesuai dengan kaidah bahasa Indonesia.

3. Mengevaluasi kinerja modul CFG menggunakan metrik accuracy, precision, recall, dan F1-score serta modul T5 menggunakan Exact Match, ROUGE-L, dan GLEU.

1.5 Manfaat Penelitian

Berdasarkan tujuan penelitian yang telah dirumuskan, manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Menyediakan alat bantu untuk mendeteksi dan mengoreksi kesalahan struktur bahasa secara otomatis guna meningkatkan efisiensi proses penyuntingan berita. Implementasi sistem ini diharapkan dapat mengurangi kesalahan kebahasaan yang terpublikasi serta mendukung peningkatan kualitas dan kredibilitas media berita daring.
2. Berpotensi membantu media massa dalam menjaga standar bahasa Indonesia yang baik dan benar sesuai dengan kaidah Tata Bahasa Baku Bahasa Indonesia, sehingga dapat mempertahankan kepercayaan pembaca dan citra profesional media.
3. Menunjukkan efektivitas pemanfaatan metode berbasis aturan (CFG) dan model deep learning (T5) dalam mendukung proses deteksi dan koreksi kesalahan struktur kalimat bahasa Indonesia.

1.6 Sistematika Penulisan

Penyusunan laporan tugas akhir ini dibagi menjadi beberapa bab untuk memberikan gambaran yang sistematis dan komprehensif mengenai alur penelitian. Adapun sistematika penulisan laporan ini adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini menguraikan latar belakang permasalahan mengenai kerentanan kesalahan struktur kalimat pada publikasi berita daring akibat tuntutan kecepatan waktu tayang. Selain itu, bab ini menjabarkan rumusan masalah, batasan permasalahan yang berfokus pada delapan kategori kesalahan sintaksis, tujuan penelitian, manfaat penelitian, dan diakhiri dengan sistematika penulisan laporan.

- Bab 2 LANDASAN TEORI
  Bab ini memuat teori-teori pokok dan konsep keilmuan yang menjadi fondasi penelitian. Pembahasan mencakup tinjauan mengenai *Natural Language Processing* (NLP), kaidah sintaksis kalimat bahasa Indonesia, arsitektur *Context Free Grammar* (CFG), metode *Part–of–Speech* (POS) Tagging, model *Text–to–Text Transfer Transformer* (T5), serta ragam metrik evaluasi yang diterapkan seperti Confusion Matrix, Exact Match, ROUGE-L, dan GLEU.

- Bab 3 METODOLOGI PENELITIAN

  Bab ini menjelaskan tahapan dan alur kerja penelitian secara sistematis. Uraian mencakup proses pengumpulan dan pra-pemrosesan data korpus Tribunnews.com, perancangan aturan berbasis CFG untuk mendeteksi kategori kesalahan (K01–K08), langkah augmentasi data untuk menyeimbangkan proporsi dataset, hingga tahapan pelatihan (finetuning) dan rancangan evaluasi untuk model IndoT5.

- Bab 4 HASIL DAN DISKUSI

  Bab ini memaparkan detail implementasi sistem beserta analisis komprehensif dari hasil evaluasi pengujian. Analisis difokuskan pada kinerja klasifikasi modul CFG dengan metrik akurasi dan F1score, kualitas generatif koreksi dari model IndoT5, pengujian prediksi kalimat, serta penjelasan terkait implementasi kedua modul ke dalam layanan REST API menggunakan FastAPI.

- Bab 5 KESIMPULAN DAN SARAN

  Bab ini menyajikan kesimpulan yang ditarik dari hasil evaluasi untuk memastikan bahwa seluruh tujuan penelitian telah tercapai. Bab ini juga memberikan saran-saran konstruktif, seperti rekomendasi penyempurnaan aturan CFG pada kategori dengan recall rendah dan perluasan kategori kesalahan, guna menjadi acuan bagi pengembangan riset selanjutnya.