

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan *Natural Language Processing* (NLP) telah mendorong pemanfaatan teknologi bahasa dalam berbagai bidang, termasuk industri media dan jurnalistik. Berbagai teknologi berbasis NLP digunakan untuk mendukung pengelolaan konten berita, mulai dari pengolahan teks, penyuntingan, hingga verifikasi informasi [1][2]. Pemanfaatan teknologi tersebut bertujuan meningkatkan efisiensi produksi berita seiring meningkatnya kebutuhan informasi dalam ekosistem digital [3]. Namun, tuntutan kecepatan publikasi juga meningkatkan risiko terjadinya *human error*, khususnya kesalahan sintaksis yang sulit dideteksi secara cepat melalui pemeriksaan manual. Oleh karena itu, diperlukan sistem pemeriksa kebahasaan otomatis yang mampu mengidentifikasi kesalahan sintaksis secara akurat pada teks berita.

Skala permasalahan ini turut tercermin dari kesenjangan antara pertumbuhan media daring dan kapasitas verifikasi redaksional formal di Indonesia. Direktorat Jenderal Informasi dan Komunikasi Publik Kementerian Komunikasi dan Informatika mencatat bahwa dari sekitar 48.000 media daring yang terdata di Indonesia, baru sekitar 400 media atau kurang dari 1% yang telah terverifikasi oleh Dewan Pers [4]. Besarnya jumlah media daring dibandingkan kapasitas verifikasi formal tersebut menunjukkan bahwa menjaga kualitas kebahasaan secara konsisten pada seluruh berita yang dipublikasikan menjadi tantangan tersendiri. Dalam kondisi demikian, kesalahan sintaksis masih berpotensi lolos hingga tahap publikasi.

Kondisi tersebut sejalan dengan temuan empiris bahwa kesalahan sintaksis memang masih muncul pada artikel berita yang telah dipublikasikan. Yusra [5] menemukan 13 kesalahan berbahasa pada tataran sintaksis hanya dari dua artikel berita pada media online Metro Jambi. Temuan serupa dilaporkan oleh Waningyun dan Ma'rifah [6], yang menemukan dua kesalahan sintaksis di antara total 14 kesalahan kebahasaan pada berbagai tataran hanya dalam satu artikel berita Suara Merdeka edisi Oktober 2024. Kedua temuan tersebut mengindikasikan bahwa kesalahan sintaksis masih dijumpai pada artikel berita daring yang telah dipublikasikan.

Signifikansi kesalahan sintaksis pada teks berita tidak dapat dinilai semata-mata berdasarkan frekuensi kemunculannya. Setiap kesalahan sintaksis muncul pada kalimat yang telah dipublikasikan dan dapat diakses oleh masyarakat luas. Kesalahan pada relasi subjek-predikat atau klausa yang menggantung, misalnya, berpotensi mengubah relasi antarinformasi dalam kalimat, menimbulkan ambiguitas makna, atau menyebabkan kesalahpahaman terhadap fakta yang disampaikan. Dalam konteks jurnalistik, kondisi tersebut dapat memengaruhi kejelasan informasi, memicu penyebaran informasi yang keliru, serta berdampak pada kredibilitas media [7]. Oleh karena itu, urgensi permasalahan ini lebih tepat dinilai dari besarnya dampak setiap kesalahan (*impact per error*) daripada semata-mata dari frekuensi kemunculannya.

Deteksi kesalahan sintaksis masih menjadi tantangan dalam pengembangan sistem pemeriksa kebahasaan otomatis. Berbeda dengan kesalahan ejaan atau kesalahan pada tingkat kata yang umumnya dapat dikenali melalui pencocokan kamus maupun pola karakter, kesalahan sintaksis melibatkan hubungan antarfrasa, klausa, dan unsur kalimat lainnya sehingga memerlukan pemahaman terhadap struktur kalimat secara lebih luas, bukan sekadar token individual [8]. Dalam konteks berita daring, kesalahan pada tataran sintaksis tidak hanya mengganggu pemahaman pembaca, tetapi juga berpotensi menurunkan kualitas penyampaian informasi dan kepercayaan publik terhadap media [9]. Oleh karena itu, diperlukan sistem deteksi yang mampu bekerja pada tingkat frasa, klausa, dan kalimat secara lebih komprehensif.

Tantangan tersebut semakin kompleks karena bahasa Indonesia memungkinkan satu kata memiliki fungsi gramatikal yang berbeda bergantung pada konteks penggunaannya. Sebagai contoh, kata *tersangka* dapat berfungsi sebagai nomina, tetapi juga dapat berfungsi sebagai verba pasif pada konteks yang berbeda. Selain itu, kesalahan sintaksis dalam bahasa Indonesia umumnya berkaitan dengan ketidaktepatan susunan frasa, relasi antarunsur kalimat, penggunaan partikel, serta penyimpangan pola kalimat dari kaidah bahasa baku [10][11]. Kompleksitas tersebut menyebabkan pendekatan berbasis aturan sederhana kurang memadai sehingga diperlukan model yang mampu memanfaatkan informasi kontekstual secara efektif. Kebutuhan tersebut menjadi salah satu tantangan utama dalam pengembangan sistem deteksi kesalahan sintaksis.

Ditemukannya kesalahan sintaksis pada artikel berita yang telah dipublikasikan dan melalui proses penyuntingan redaksional juga menunjukkan adanya keterbatasan pemeriksaan manual, sebagaimana dilaporkan pada penelitian

terhadap media daring lain [5][12]. Keterbatasan tersebut bukan semata-mata mencerminkan kelalaian individu penyunting, melainkan merupakan konsekuensi dari proses editorial media daring yang berorientasi pada kecepatan publikasi. Akibatnya, efektivitas pemeriksaan sintaksis sangat bergantung pada waktu yang tersedia, beban kerja penyunting, serta kompleksitas struktur kalimat yang diperiksa. Tantangan tersebut menjadi semakin besar mengingat jumlah media daring di Indonesia telah mencapai puluhan ribu sebagaimana telah dijelaskan sebelumnya. Kesalahan yang lolos hingga tahap publikasi juga terbukti dapat memengaruhi dimensi *accuracy* dalam kerangka kredibilitas media; Abanat dkk. [7] menemukan bahwa 50% berita yang dianalisis tidak memenuhi dimensi tersebut akibat kesalahan penulisan, meskipun telah melalui proses redaksional. Selain *accuracy*, penilaian kredibilitas media daring juga mempertimbangkan dimensi *believability*, *bias*, dan *completeness* [7]. Kondisi tersebut menunjukkan perlunya mekanisme deteksi otomatis sebagai lapisan pemeriksaan tambahan yang mampu bekerja secara konsisten pada skala volume produksi berita yang besar.

Dengan demikian, urgensi penelitian ini bertumpu pada tiga aspek utama, yaitu besarnya dampak setiap kesalahan terhadap kejelasan informasi dan kredibilitas media, keterbatasan pemeriksaan manual di tengah skala produksi berita yang besar, serta kebutuhan akan sistem yang mampu mendeteksi seluruh kategori kesalahan sintaksis secara komprehensif. Dengan kata lain, urgensi penelitian ini tidak ditentukan semata-mata oleh frekuensi kemunculan kesalahan, melainkan oleh konsekuensi yang ditimbulkan ketika kesalahan tersebut terjadi.

Penelitian terkait deteksi dan koreksi kesalahan kebahasaan bahasa Indonesia telah berkembang dalam beberapa tahun terakhir, meliputi deteksi kesalahan ejaan [13], perbaikan kesalahan ketik [14], deteksi kata baku [15], serta koreksi gramatikal berbasis *Transformer* [16]. Di Universitas Multimedia Nusantara, pengembangan sistem deteksi kesalahan kebahasaan juga dilakukan melalui platform U-Tapis yang dirancang sebagai sistem modular untuk berbagai jenis kesalahan bahasa Indonesia [17]. Meskipun demikian, penelitian yang secara khusus berfokus pada deteksi kesalahan sintaksis berbasis frasa masih relatif terbatas. Di sisi lain, penerapan *Conditional Random Field* (CRF) telah menunjukkan hasil yang menjanjikan pada tugas *part-of-speech tagging* bahasa Indonesia [18], namun pemanfaatannya secara spesifik untuk deteksi kesalahan sintaksis berbasis frasa pada domain berita daring masih belum banyak dieksplorasi.

Perkembangan terkini menunjukkan bahwa integrasi model deteksi sekuensial dengan pendekatan koreksi berbasis *deep learning*, khususnya arsitektur

Text-to-Text Transfer Transformer (T5), merupakan pendekatan yang menjanjikan untuk membangun sistem deteksi dan koreksi yang lebih adaptif [8][19]. Ketersediaan korpus sebanyak 10.000 artikel Tribunnews yang mencakup lebih dari 160.000 kalimat, yang telah dikumpulkan oleh tim U-Tapis, membuka peluang untuk mengembangkan model deteksi berbasis CRF yang terintegrasi dengan mekanisme koreksi berbasis T5 pada domain berita daring Indonesia [10].

Berdasarkan uraian tersebut, masih terdapat kesenjangan penelitian berupa terbatasnya penelitian yang secara khusus berfokus pada deteksi kesalahan sintaksis berbasis frasa yang terintegrasi dengan mekanisme koreksi otomatis pada domain berita daring Indonesia. Sebagian besar penelitian sebelumnya berfokus pada deteksi atau koreksi secara terpisah serta lebih banyak menitikberatkan pada kesalahan ejaan, tata bahasa umum, atau kesalahan pada tingkat kata. Oleh karena itu, diperlukan pendekatan yang tidak hanya mampu mendeteksi lokasi kesalahan sintaksis secara akurat, tetapi juga menghasilkan koreksi yang sesuai dengan konteks kalimat. Dalam konteks tersebut, Conditional Random Field (CRF) dipilih sebagai metode deteksi karena mampu memodelkan dependensi antarlabel pada proses pelabelan sekuens sehingga dapat mengidentifikasi batas-batas frasa yang menyimpang sebagai satu kesatuan unit sintaksis. Hasil deteksi tersebut kemudian digunakan sebagai masukan bagi IndoT5 untuk menghasilkan rekomendasi perbaikan kalimat secara kontekstual.

Penelitian ini mengusulkan sistem deteksi dan koreksi kesalahan sintaksis bahasa Indonesia yang mengintegrasikan CRF dan IndoT5. CRF digunakan untuk mengidentifikasi lokasi serta kategori kesalahan sintaksis berbasis frasa, sedangkan IndoT5 digunakan untuk menghasilkan rekomendasi perbaikan kalimat secara kontekstual berdasarkan hasil deteksi tersebut. Melalui pendekatan ini, penelitian diharapkan dapat berkontribusi terhadap pengembangan sistem pemeriksa sintaksis bahasa Indonesia yang tidak hanya mampu mengidentifikasi kesalahan, tetapi juga menghasilkan koreksi yang sesuai dengan konteks kalimat.

1.2 Rumusan Masalah

Rumusan masalah penelitian ini dijabarkan dalam pertanyaan penelitian sebagai berikut:

1. Bagaimana pendekatan *Conditional Random Field* (CRF) berbasis frasa dapat digunakan untuk mendeteksi kesalahan sintaksis pada teks berita daring berbahasa Indonesia?

2. Bagaimana model *Text-to-Text Transfer Transformer* (T5) dapat digunakan untuk melakukan koreksi otomatis terhadap kesalahan sintaksis yang terdeteksi?
3. Bagaimana kinerja sistem deteksi dan koreksi kesalahan sintaksis yang dibangun pada teks berita daring berbahasa Indonesia?

1.3 Batasan Masalah

Batasan permasalahan dalam penelitian ini ditetapkan untuk membatasi ruang lingkup pembahasan agar tetap fokus dan terarah. Adapun batasan permasalahan dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini difokuskan pada deteksi dan koreksi empat kategori utama yang dijabarkan menjadi 12 aturan kesalahan sintaksis berbasis relasi antarfrasa pada teks berita daring berbahasa Indonesia.
2. Jenis kesalahan sintaksis yang dibahas dibatasi pada:
 - (a) **Kesalahan Relasi Subjek-Predikat:** Fokus pada identifikasi ketidaklengkapan unsur inti kalimat dan gangguan hubungan sintaksis primer.
 - (b) **Kesalahan Klausa Subordinatif:** Fokus pada identifikasi kalimat dependen yang tidak memiliki induk kalimat (independent clause).
 - (c) **Redundansi Konjungsi Subordinatif:** Fokus pada identifikasi penggunaan ganda konjungsi yang memiliki fungsi logika serupa dalam satu konstruksi kalimat.
 - (d) **Ketidaktepatan Konjungsi Idiomatik dan Korelatif:** Fokus pada kesalahan penggunaan pasangan kata yang sudah baku (idiomatik) dan konjungsi berpasangan (korelatif).
3. Penelitian ini berfokus pada deteksi dan koreksi kesalahan sintaksis berbasis frasa. Perubahan pada aspek kebahasaan lain, seperti ejaan, morfologi, semantik, dan pragmatik, yang mungkin muncul sebagai efek samping selama proses koreksi tidak dianalisis secara khusus dan tidak menjadi objek evaluasi utama penelitian. Evaluasi penelitian difokuskan pada kemampuan sistem dalam mendeteksi dan mengoreksi kesalahan sintaksis sesuai kategori yang telah ditetapkan.

4. Penelitian ini menggunakan *Conditional Random Field* (CRF) untuk tahap deteksi kesalahan sintaksis dan *Text-to-Text Transfer Transformer* (T5) untuk tahap koreksi.
5. Data penelitian berasal dari korpus 10.000 artikel berita daring berbahasa Indonesia yang bersumber dari Tribunnews serta augmentasi data untuk mendukung pelatihan dan evaluasi model.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model *Conditional Random Field* (CRF) berbasis frasa untuk mendeteksi kesalahan sintaksis pada teks berita daring berbahasa Indonesia.
2. Mengembangkan mekanisme koreksi otomatis menggunakan arsitektur *Text-to-Text Transfer Transformer* (T5) berdasarkan hasil deteksi kesalahan sintaksis.
3. Mengevaluasi kinerja sistem deteksi dan koreksi kesalahan sintaksis yang dihasilkan.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut.

1.5.1 Manfaat Teoretis

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode deteksi dan koreksi kesalahan sintaksis bahasa Indonesia berbasis frasa menggunakan pendekatan *Conditional Random Field* (CRF) dan *Text-to-Text Transfer Transformer* (T5). Selain itu, penelitian ini dapat memperkaya kajian *Natural Language Processing* (NLP) untuk bahasa Indonesia, khususnya pada bidang deteksi dan koreksi kesalahan sintaksis pada teks berita daring.

1.5.2 Manfaat Praktis

1. Menghasilkan sistem deteksi dan koreksi kesalahan sintaksis bahasa Indonesia berbasis frasa yang mampu membantu proses pemeriksaan kebahasaan pada teks berita daring.
2. Menjadi modul pendukung dalam pengembangan platform U-Tapis sebagai sistem pemeriksa kebahasaan berbasis kecerdasan buatan.
3. Membantu wartawan dan editor dalam meningkatkan efisiensi proses penyuntingan serta kualitas kebahasaan teks berita melalui deteksi dan koreksi otomatis.
4. Menjadi landasan pengembangan penelitian lanjutan pada bidang teknologi pemeriksa tata bahasa bahasa Indonesia berbasis integrasi model sekuensial dan generatif.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun dalam lima bab dengan uraian sebagai berikut.

- Bab 1 **PENDAHULUAN**

Membahas latar belakang penelitian, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan skripsi.

- Bab 2 **LANDASAN TEORI**

Berisi kajian teori dan penelitian terdahulu yang mendukung penelitian ini. Pembahasan meliputi konsep sintaksis bahasa Indonesia, algoritma *Conditional Random Field* (CRF), arsitektur *Text-to-Text Transfer Transformer* (T5), serta metrik evaluasi kinerja sistem seperti *precision*, *recall*, *accuracy*, dan *F1-score*.

- Bab 3 **METODOLOGI PENELITIAN**

Menjelaskan tahapan penelitian yang dilakukan, meliputi proses pengumpulan dan penyusunan korpus data, prapemrosesan dan pelabelan data berbasis frasa, perancangan serta pelatihan model *Conditional Random Field* (CRF) untuk deteksi kesalahan sintaksis, perancangan

dan implementasi model koreksi otomatis berbasis arsitektur *Text-to-Text Transfer Transformer* (T5), serta metode evaluasi kinerja sistem secara terpisah maupun terintegrasi. Bab ini juga menjelaskan implementasi sistem yang digunakan untuk proses inferensi dan pengujian model.

- Bab 4 **HASIL DAN DISKUSI**

Menyajikan hasil implementasi sistem serta hasil pengujian model deteksi berbasis CRF dan model koreksi berbasis T5. Pada bab ini juga dilakukan analisis terhadap hasil evaluasi dan uji coba sistem untuk mengidentifikasi kelebihan serta keterbatasan pendekatan yang diusulkan.

- Bab 5 **KESIMPULAN DAN SARAN**

Berisi simpulan yang diperoleh dari hasil penelitian serta saran untuk pengembangan sistem dan penelitian lanjutan yang berkaitan dengan deteksi dan koreksi kesalahan sintaksis bahasa Indonesia.

