BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Di era digital yang serba cepat, arus informasi melalui media massa daring menuntut tim jurnalistik untuk memproduksi konten di bawah tenggat waktu yang ketat. Tuntutan kecepatan ini sering kali berdampak pada kualitas bahasa, yang mengakibatkan tingginya frekuensi kesalahan linguistik pada berita yang dipublikasikan. Kesalahan berbahasa ini mencakup berbagai aspek, mulai dari ejaan, pemilihan diksi, hingga struktur kalimat yang tidak sesuai dengan kaidah Bahasa Indonesia yang baik dan benar. Fenomena ini tidak hanya terjadi di media profesional, tetapi juga meluas ke platform berita media sosial seperti LINE TODAY, di mana kesalahan pada ejaan, morfologi, dan sintaksis juga kerap ditemukan [1]. Meskipun penggunaan Bahasa Indonesia yang standar telah diatur dalam Peraturan Presiden Nomor 63 Tahun 2019, implementasinya di lapangan masih lemah, terbukti dengan maraknya kesalahan ejaan pada artikel opini di media massa [2].

Menyadari tantangan tersebut, Universitas Multimedia Nusantara menginisiasi pengembangan platform inovatif berbasis kecerdasan buatan (AI) bernama U-Tapis [3]. Dikembangkan melalui kemitraan dengan industri media seperti PT Tribun Digital Online, U-Tapis berfungsi sebagai ekosistem yang terus berkembang untuk meningkatkan keterampilan berkomunikasi [3]. Berbagai modul spesifik telah dikembangkan untuk mengatasi masalah tertentu, seperti deteksi ”kata luluh” (*melting word*) [4] dan deteksi pelecehan (*harassment detection*) [5].

Namun, analisis terhadap penelitian sebelumnya menunjukkan adanya celah fungsionalitas yang signifikan pada U-Tapis. Sebuah penelitian versi awal secara eksplisit menyebutkan bahwa sistem tersebut masih gagal memperbaiki kesalahan pengetikan pada kata asing yang seharusnya dicetak miring [3]. Aturan penggunaan huruf miring dan tegak ini merupakan bagian fundamental dari Pedoman Umum Ejaan Bahasa Indonesia (PUEBI) yang sering diabaikan. Urgensi penerapan aturan ini terletak pada fungsi huruf miring sebagai indikator visual utama untuk membedakan istilah asing dari kosakata baku Bahasa Indonesia. Tidak digunakannya penanda tersebut dapat memicu kesalahan interpretasi, di mana kata

asing yang ditulis tegak berpotensi dianggap sebagai kesalahan penulisan (*typo*). Sebagai contoh, kata ”issue” yang ditulis tegak dapat disalahartikan pembaca sebagai kesalahan pengetikan dari kata baku ”isu”. Ambiguitas visual semacam ini tidak hanya menghambat pemahaman konten, tetapi juga mengindikasikan rendahnya akurasi penyuntingan yang berdampak pada kredibilitas berita. Berbagai analisis mengonfirmasi bahwa kesalahan penggunaan huruf miring masih sangat sering terjadi pada artikel di media cetak maupun daring [6].

Oleh karena itu, penelitian ini bertujuan untuk mengisi kekosongan tersebut dengan mengembangkan sebuah model deteksi. Pendekatan yang diusulkan adalah metode hibrida yang menggabungkan *Rule-Based* untuk menangani aturan PUEBI yang pasti, dan *Conditional Random Field* (CRF) untuk menangani kasus yang memerlukan pemahaman konteks. Pendekatan *Rule-Based* terbukti dapat diimplementasikan untuk mendeteksi kesalahan gramatikal berdasarkan aturan PUEBI, seperti pada kasus huruf kapital [7]. Sementara itu, CRF merupakan model statistik yang sangat efektif untuk tugas pelabelan data sekuensial (*sequential labeling*), karena kemampuannya mempertimbangkan konteks kata di sekitarnya untuk menentukan label yang paling mungkin [8].

Sebagai konteks, dalam perkembangan terkini di bidang *Natural Language Processing* (NLP), arsitektur *Deep Learning* yang kompleks sering digunakan untuk tugas pelabelan sekuensial. Salah satunya adalah *Long Short-Term Memory* (LSTM), sebuah jenis Jaringan Saraf Berulang yang efektif untuk tugas seperti koreksi ejaan karena kemampuannya mengingat informasi jangka panjang [9]. Arsitektur ini kemudian dikembangkan menjadi *Bidirectional* LSTM (BiLSTM), yang mampu memproses data dari dua arah untuk mendapatkan pemahaman konteks yang lebih kaya. Kombinasi BiLSTM dengan CRF (BiLSTM-CRF) kini menjadi salah satu pendekatan *state-of-the-art* untuk tugas *Named Entity Recognition* (NER) dalam Bahasa Indonesia [10] [11].

Meskipun arsitektur *Deep Learning* mendominasi performa terkini, pendekatan statistik tetap memiliki relevansi yang tinggi, terutama dalam menangani bahasa dengan karakteristik morfologi yang kaya [12]. Studi literatur menunjukkan bahwa pada kondisi linguistik tersebut, algoritma statistik mampu mencapai akurasi identifikasi dan klasifikasi entitas yang optimal [12]. Selain itu, model CRF secara spesifik menawarkan keunggulan dalam pembelajaran berbasis konteks (*context dependent learning*) yang vital untuk menyelesaikan dependensi antarlabel [12]. Berlandaskan kapabilitas teknis tersebut dalam menangani kompleksitas bahasa, penelitian ini mengadopsi CRF sebagai model utama yang

didukung oleh pendekatan *Rule-Based* dalam skema hibrida.

1.2 Rumusan Masalah

Sesuai dengan latar belakang masalah yang telah dipaparkan, maka dapat dirumuskan masalah sebagai berikut:

1. Bagaimana mengimplementasikan model hibrida yaitu *Rule-Based* dan *Conditional Random Field* (CRF) untuk membangun model yang mampu mendeteksi kesalahan penggunaan huruf miring dan huruf tegak disertai tanda kutip pada teks berita sesuai kaidah PUEBI?
2. Seberapa tinggi tingkat akurasi dan performa model hibrida yang dikembangkan dalam mendeteksi kata atau frasa yang penggunaannya tidak sesuai aturan pada dataset teks berita?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, adapun tujuan yang hendak dicapai dari penelitian ini adalah:

1. Membangun sebuah model hibrida fungsional yang menggunakan *Rule-Based* dan *Conditional Random Field* (CRF) untuk mendeteksi secara otomatis kesalahan penggunaan huruf miring dan tegak disertai tanda kutip.
2. Mengevaluasi performa model yang dikembangkan secara kuantitatif menggunakan metrik standar (*accuracy, precision, recall, F1-score*) untuk mengukur efektivitasnya sebagai modul deteksi baru.

1.4 Batasan Masalah

1. Sistem yang dikembangkan hanya sebatas mendeteksi kata atau frasa yang berpotensi salah dan tidak memberikan rekomendasi koreksi.
2. Penelitian ini berfokus pada implementasi pendekatan hibrida *Rule-Based* dan *Conditional Random Field* (CRF), dan tidak melakukan perbandingan komprehensif dengan arsitektur *deep learning* lain seperti BiLSTM.
3. Cakupan deteksi terbatas hanya pada kesalahan penggunaan huruf miring sesuai aturan PUEBI.

4. Dataset yang digunakan adalah korpus artikel berita daring berbahasa Indonesia yang dianotasi untuk tugas deteksi selama periode penelitian.
5. Model tidak selalu bisa mendeteksi kesalahan yang ada pada kalimat secara akurat.
6. Beberapa kategori seperti Bahasa Asing, Bahasa Daerah, dan Nama Ilmiah memiliki keterbatasan dikarenakan jumlah data kamus yang terbatas.

1.5 Urgensi Penelitian

Urgensi penelitian ini didasari oleh adanya celah fungsionalitas yang teridentifikasi pada platform U-Tapis, yaitu kegagalan dalam menangani kesalahan penggunaan huruf miring. Mengingat U-Tapis dirancang untuk jurnalis, menyediakan alat yang mampu mendeteksi dan menyoroti potensi kesalahan fundamental PUEBI menjadi langkah krusial pertama untuk membantu mereka menjaga kualitas tulisan.

1.6 Luaran Penelitian

1. Model *website* U-Tapis untuk mendeteksi penggunaan huruf miring.
2. Artikel ilmiah terakreditasi Sinta.

1.7 Manfaat Penelitian

Adapun manfaat penelitian ini adalah:

1. Memberikan kontribusi praktis dengan menyempurnakan platform U-Tapis melalui penambahan modul fungsional baru, sehingga secara langsung meningkatkan kualitas linguistik konten yang diproduksi oleh jurnalis, editor, dan mahasiswa.
2. Memberikan kontribusi teoretis pada bidang Natural Language Processing (NLP) untuk Bahasa Indonesia, khususnya dalam menyediakan solusi untuk tugas koreksi ejaan kontekstual yang sebelumnya belum teratasi secara spesifik.