BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kesalahan penulisan dalam bahasa Indonesia memiliki dampak yang signifikan terhadap pemahaman pembaca dan kredibilitas tulisan [1]. Penelitian terhadap media massa menunjukkan bahwa kesalahan berbahasa seperti ejaan, morfologi, sintaksis, dan sematik dapat merusak kredibilitas pesan serta menghalangi masyarakat dalam memahami informasi yang disampaikan [2]. Penggunaan partikel, preposisi, dan diksi yang tidak tepat dalam konteks kalimat juga dapat menimbulkan ambiguitas makna sehingga menghambat efektivitas komunikasi ilmiah dan publik. Sejalan dengan permasalahan tersebut, penelitian sebelumnya menunjukkan bahwa deteksi otomatis kesalahan gramatikal mampu meningkatkan keterbacaan dan kepercayaan pembaca terhadap teks [3].

Natural Language Processing (NLP) merupakan ilmu yang mempelajari komunikasi manusia dengan komputer melalui bahasa alami [4][5]. Tujuan utamanya adalah membangun model komputasi agar komputer dapat memahami, memanipulasi, dan melakukan tugas berdasarkan bahasa alami [6], mulai dari terjemahan mesin hingga asisten pribadi canggih dengan pemahaman konteks pengguna [7]. Salah satu penerapan dasar NLP adalah analisis dan peningkatan kualitas teks yang mendukung proses belajar dan komunikasi [8]. Berbagai jenis kesalahan seperti ejaan, tipografi, sintaks, diksi, serta penggunaan kosakata asing perlu diidentifikasi karena melemahkan makna dan keterbacaan [9][10][11]. Oleh karena itu, sistem deteksi kesalahan gramatikal menjadi salah satu cabang penelitian penting, yang melibatkan analisis leksikal, sintaksis, dan semantik untuk mendeteksi kesalahan penulisan teks secara otomatis [12].

Penelitian untuk mendeteksi kesalahan penulisan teks di Indonesia telah menunjukkan kemajuan yang nyata [13]. Sejumlah studi berhasil mendeteksi kesalahan ejaan dan salah ketik dengan metode seperti Damerau-Levenshtein Distance [14] dan Random Forest [15], serta mengidentifikasi kesalahan penulisan frasa preporsisi yang seharusnya dipisah namun tertulis menyatu [16]. Sejumlah pendekatan tersebut umumnya berfokus pada kesalahan ejaan dan pemisahan kata. Di sisi lain, penelitian terkini mulai memperluas cakupan deteksi dalam analisis gramatikal dan struktur kalimat, salah satunya adalah U-Tapis. Dalam konteks

ini, U-Tapis ikut berkontribusi dalam mengembangkan sistem deteksi dan koreksi kesalahan gramatikal bagi teks jurnalistik [17]. Meskipun demikian, penggunaan partikel masih belum mendapatkan perhatian penelitian yang setara dengan jenis kesalahan lain [18].

Kesalahan penggunaan partikel dalam bahasa Indonesia seperti *lah*, *kah*, *pun*, dan *per* memiliki karakteristik yang kompleks dan sangat bergantung pada aturan-aturan bahasa serta konteks kalimat [19]. Kompleksitas tersebut menuntut penerapan metode klasifikasi yang mampu mempelajari dan menangani data teks dengan variasi fitur yang tinggi. Oleh karena itu, penelitian ini memilih algoritma XGBoost yang dikenal unggul dalam menganalisis data teks dan memberikan performa yang lebih baik dibandingkan metode seperti Naïve Bayes maupun Random Forest [20][21]. Keunggulan XGBoost dalam menangani data teks juga telah dibuktikan pada berbagai penelitian sebelumnya pada teks berbahasa Indonesia, termasuk pada tugas analisis sentimen dan klasifikasi berita, yang menunjukkan performa kompetitif dan stabil [22][23][24].

Penelitian ini bertujuan untuk membangun model deteksi kesalahan partikel pada teks berbahasa Indonesia menggunakan kombinasi XGBoost dan *rule-based system*. Fokus utama penelitian ini adalah menguji kemampuan kombinasi XGBoost dengan *rule-based system* dalam mendeteksi kesalahan partikel pada teks berbahasa Indonesia. Penelitian ini diharapkan dapat memberikan kontribusi berupa model deteksi yang andal, sekaligus memperkaya literatur NLP untuk Bahasa Indonesia, khususnya pada aspek gramatikal.

1.2 Rumusan Masalah

Rumusan masalah yang dirumuskan berdasarkan latar belakang di atas adalah sebagai berikut:

1. Bagaimana cara mendeteksi kesalahan penggunaan partikel Bahasa Indonesia menggunakan kombinasi algoritma XGBoost dan *rule-based system*?
2. Bagaimana performa model hasil implementasi tersebut dalam mendeteksi kesalahan partikel berdasarkan hasil evaluasi terhadap data uji?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model deteksi kesalahan penulisan partikel Bahasa Indonesia menggunakan kombinasi algoritma XGBoost dan *rule-based system* sesuai kaidah Ejaan Yang Disempurnakan (EYD).
2. Mengevaluasi kinerja model dalam mendeteksi kesalahan partikel berdasarkan metrik akurasi, presisi, recall, dan F1-Score pada dataset Bahasa Indonesia.

1.4 Urgensi Penelitian

Penelitian ini memiliki urgensi tinggi karena kesalahan penulisan partikel seperti *pun*, *per*, *-lah*, dan *-kah* masih sering ditemukan bahkan pada teks jurnalistik dan akademik yang seharusnya berbahasa baku. Meskipun telah banyak penelitian yang berfokus pada kesalahan ejaan dan pemisahan kata, penelitian terkait deteksi kesalahan partikel dalam Bahasa Indonesia masih terbatas. Oleh karena itu, penelitian ini penting untuk:

- Mengisi celah penelitian pada deteksi kesalahan partikel Bahasa Indonesia.
- Meningkatkan kualitas pemeriksaan bahasa otomatis melalui integrasi pendekatan pembelajaran mesin dan aturan linguistik.
- Mendukung pengembangan teknologi bahasa Indonesia pada ranah NLP.

1.5 Luaran Penelitian

Luaran yang dihasilkan dari penelitian ini meliputi:

1. Sebuah API aplikasi web U-Tapis yang berfungsi untuk mendeteksi kesalahan penggunaan partikel sesuai kaidah EYD, sehingga dapat diintegrasikan dengan sistem pemeriksa ejaan maupun aplikasi jurnalistik lainnya.
2. Artikel ilmiah yang memuat hasil penelitian ini dan direncanakan untuk dipublikasikan pada jurnal nasional terakreditasi SINTA.

1.6 Manfaat Penelitian

Penelitian ini diharapkan memberikan beberapa manfaat yang relevan dengan pengembangan sistem pemeriksa ejaan Bahasa Indonesia. Pertama,

penelitian ini dapat meningkatkan akurasi sistem pemeriksa ejaan, khususnya dalam mendeteksi kesalahan penulisan partikel yang sebelumnya belum terakomodasi dalam sistem U-Tapis. Kedua, hasil penelitian ini berpotensi mendukung peningkatan kualitas penulisan jurnalistik dengan membantu jurnalis menjaga ketepatan penggunaan ejaan sesuai kaidah bahasa Indonesia. Selain itu, penelitian ini turut memperluas penerapan Natural Language Processing (NLP) dan kecerdasan buatan dalam bidang kebahasaan, terutama dalam konteks bahasa Indonesia yang memiliki aturan ejaan kompleks. Lebih jauh, penelitian ini juga dapat menjadi dasar bagi pengembangan penelitian lanjutan yang berfokus pada sistem pemeriksa ejaan otomatis maupun analisis sintaksis berbasis pembelajaran mesin. Dengan demikian, penelitian ini berkontribusi dalam meningkatkan efisiensi proses penyuntingan teks sekaligus memperkuat pemanfaatan teknologi AI dalam dunia jurnalistik dan pendidikan.