

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tulisan merupakan media informasi dan komunikasi yang sudah ada semenjak dahulu, bahkan sebelum masehi. Pada saat itu, manusia cuma dapat berkomunikasi melalui gambar atau hieroglif dan simbol yang ada benda-benda alam seperti kayu, batu, dan dinding gua, alam seperti kayu, batu, dan dinding gua. Seiring dengan perkembangan teknologi, kegiatan menulis yang telah beralih dari penggunaan kertas ke media digital memberikan keuntungan dalam efisiensi dan aksesibilitas informasi. Meskipun kemajuan teknologi telah memudahkan kegiatan menulis tidak dapat dipungkiri bahwa kesalahan penulisan seperti typo masih sering terjadi (Faizan et al., n.d.). Perkembangan teknologi saat ini telah dimanfaatkan dalam berbagai aspek kehidupan manusia, seperti komunikasi, visualisasi, pendidikan, bisnis, dan berbagai kebutuhan lainnya (Oktavia Suzanti et al., 2024a). Penulisan karya ilmiah, seperti jurnal penelitian, harus menggunakan bahasa baku yang sesuai dengan kaidah Ejaan Yang Disempurnakan (EYD), termasuk penggunaan kosakata standar yang mengacu pada Kamus Besar Bahasa Indonesia (KBBI). (Risang Agung Samudero et al., 2024).

Menurut Kamus Besar Bahasa Indonesia (KBBI), bahasa Indonesia dibedakan menjadi dua jenis, yaitu bahasa baku dan bahasa tidak baku. Bahasa baku merupakan ragam bahasa yang telah ditetapkan sebagai standar penggunaan bahasa yang baik dan benar dalam situasi resmi, baik secara lisan maupun tulisan. Sementara itu, bahasa tidak baku adalah bentuk bahasa yang digunakan dalam komunikasi tidak resmi dengan penggunaan kosakata, struktur kalimat, ejaan, maupun pelafalan yang tidak selalu sesuai dengan kaidah bahasa yang berlaku. Oleh karena itu, diperlukan adanya sistem atau metode yang mampu memperbaiki kesalahan penulisan secara otomatis agar kualitas informasi yang disampaikan tetap akurat dan mudah dipahami oleh pembaca (Qulub et al., 2023).

Natural Language Processing (NLP) merupakan salah satu komponen dari Text Mining yang berfokus pada hubungan antara komputer dan bahasa alami manusia. Tujuan dari NLP adalah untuk mendesain dan membangun perangkat lunak yang dapat menganalisis, memahami, dan menghasilkan bahasa manusia

yang dapat digunakan secara alami, sehingga saat berkomunikasi dengan komputer seakan akan berkomunikasi dengan manusia lain (Maghfira et al., 2017a). Salah satu contoh aplikasi NLP adalah *spelling correction* atau perbaikan ejaan, yang membantu mendeteksi dan memperbaiki kesalahan dalam penulisan.

Spelling correction adalah proses mendeteksi dan memberikan saran untuk kata-kata yang salah eja di dalam suatu teks, baik dengan cara mengganti kata yang salah dengan bentuk yang benar maupun dengan memberikan beberapa alternatif perbaikan yang paling mendekati kata yang dimaksud. Sedangkan *Spelling Corrector* adalah fitur koreksi otomatis untuk melakukan proses deteksi kesalahan kata dan pemberian saran kata yang mengalami kesalahan eja pada suatu teks dan *spelling corrector* juga merupakan fitur atau aplikasi yang akan melakukan proses tersebut. Fitur ini mencari kata-kata yang salah berdasarkan acuan yaitu Kamus Besar Bahasa Indonesia (KBBI) yang digunakan aplikasi. Sedangkan saran kata diberikan dengan perhitungan algoritma yang juga digunakan oleh aplikasi. Sistem *spelling correction* tidak hanya mampu mengenali kesalahan pada bahasa formal, tetapi juga pada bahasa tidak baku atau *informal* yang sering digunakan dalam media sosial dan komunikasi daring.

Kesalahan penulisan kata umumnya berkaitan dengan ketidaktepatan penggunaan atau penempatan kata dalam suatu kalimat sehingga dapat menimbulkan kesalahan ejaan dan mengubah makna kata tersebut. Dalam bahasa Indonesia, misalnya penulisan kata “sya” yang seharusnya ditulis “saya”, dapat menyebabkan kata menjadi tidak bermakna atau menimbulkan pemahaman yang berbeda. Selain itu, kesalahan ejaan seperti ini juga dapat membuat pembaca kesulitan dalam memahami maksud serta menentukan bentuk kata atau kalimat yang benar (Risang Agung Samudero et al., 2024). Pemeriksa ejaan umumnya dibedakan menjadi dua kategori, yaitu pemeriksa kesalahan non-kata dan pemeriksa kesalahan kata nyata. Pemeriksa kesalahan non-kata bertugas mendeteksi kata yang salah eja akibat kesalahan pengetikan atau tipografi. Sementara itu, pemeriksa kesalahan kata nyata berfokus pada identifikasi penggunaan atau penempatan kata yang tidak sesuai dalam suatu kalimat. Kesalahan penulisan kata atau biasa disebut *typographical error* adalah salah satu kesalahan yang sering dilakukan oleh penulis saat menulis suatu dokumen teks

Bahasa Indonesia. Kesalahan penulisan kata atau *typographical error* dapat dilakukan saat proses pengetikan kata. Kesalahan tersebut merupakan kesalahan yang terjadi pada pengetikan disebabkan karena kelalaian penulis seperti salah huruf atau *slip* pada keyboard yang berdekatan.

Fenomena kesalahan ejaan dan penulisan kata (*spelling correction*) merupakan salah satu dampak dari kebiasaan berkomunikasi secara cepat di media sosial. Pengguna sering kali tidak memperhatikan kaidah penulisan yang benar, seperti penggunaan huruf kapital, penulisan kata baku, atau kesalahan ketik akibat pengetikan yang terburu-buru. Misalnya, kata “sya” untuk “saya”, “bgt” untuk “banget”, “gk” untuk “nggak”, atau “km” untuk “kamu” menjadi bentuk yang umum digunakan, meskipun tidak sesuai dengan ejaan yang benar. Selain itu, kesalahan ejaan juga dapat terjadi pada bentuk kata yang seharusnya dipisah, seperti “dirumah” yang seharusnya “di rumah”, atau kata yang salah ketik seperti “rumahskt” yang seharusnya “rumah sakit”.(Nababan et al., 2024). Kesalahan kesalahan ini meskipun tampak sepele dapat mengurangi kredibilitas dan profesionalisme(Faizan et al., n.d.). Proses pengecekan kesalahan pengetikan dengan cara manual akan memakan banyak waktu dan memerlukan suatu sumber pasti sebagai acuan bahwa kata tersebut memang salah dalam proses penulisannya(Risang Agung Samudero et al., 2024). Oleh karena itu, terdapat satu topik untuk mengatasi masalah kesalahan ejaan dan kesalahan penulisan dalam pencarian informasi menggunakan query, yaitu pembetulan ejaan. Pembetulan kesalahan ejaan merupakan fitur pembetulan otomatis yang digunakan untuk melakukan proses pendeteksian dalam kesalahan ejaan dan penulisan kata dan juga memberikan saran kata dalam kesalahan ejaan dan penulisan pada teks(Oktavia Suzanti et al., 2024a). Untuk mendeteksi kesalahan penulisan maupun ejaan, diperlukan sebuah aplikasi yang dikenal sebagai *spelling checker*. Aplikasi ini bekerja dengan memeriksa setiap kata dalam teks untuk menemukan kesalahan penulisan, kemudian memberikan rekomendasi berupa saran perbaikan kata yang dianggap paling sesuai.

Pada tahun 2020 , penelitian perbaikan ejaan kata dilakukan oleh Aji Prasetya Wibawa dkk. Algoritma yang digunakan yaitu *Damerau Levenshtein Distance* yang disandingkan dengan metode *Empiris*, yang mana pada penelitian

tersebut dataset yang digunakan dalam penelitian tersebut berasal dari dua cerita dongeng yang diambil dari situs ceritadongengrakryat.com. Data yang digunakan berjumlah 1266 kata yang mengandung kesalahan ejaan sehingga memerlukan proses koreksi. Kata-kata dianggap salah apabila tidak ditemukan dalam Kamus Besar Bahasa Indonesia (KBBI), meskipun terdapat dalam teks dongeng tersebut.). Pada penelitian lainnya juga Algoritma *Damerau Levenshtein Distance* banyak diimplementasikan pada suatu aplikasi Identifikasi ejaan dan kesalahan kata dalam Bahasa Indonesia berbasis web. Korpus yang digunakan berasal dari kosakata Kamus Besar Bahasa Indonesia (KBBI). Banyak peneliti yang terdorong untuk melakukan penelitian guna menentukan solusi permasalahan dengan menyusun aplikasi automatic spell checker (Maghfira et al., 2017a).

Damerau Levenshtein Distance merupakan sebuah algoritma yang digunakan untuk menghitung tingkat kemiripan antara dua kata berdasarkan jumlah operasi penyuntingan (*edit*) yang diperlukan untuk mengubah satu kata menjadi kata lain. Operasi yang digunakan dalam *Damerau Levenshtein Distance* meliputi penyisipan (*insertion*), penghapusan (*deletion*), penggantian (*substitution*) dan pertukaran (*transposition*) karakter. Levenshtein Distance merupakan kasus khusus dari algoritma *Edit Distance*, di mana setiap operasi memiliki bobot (*weight*) yang sama, yaitu 1 satuan (Oktavia Suzanti et al., 2024a). Metode *Damerau Levenshtein Distance* dapat dimanfaatkan untuk memperbaiki serta menormalisasi kata yang mengalami kesalahan ejaan menjadi bentuk kata baku. Dengan metode ini, kata-kata tidak baku yang ditulis pengguna dapat dikoreksi agar sesuai dengan kosakata yang terdapat dalam Kamus Besar Bahasa Indonesia (KBBI). Penerapan metode tersebut diharapkan mampu meningkatkan tingkat akurasi sistem koreksi ejaan dibandingkan penelitian-penelitian sebelumnya. (Oktavia Suzanti et al., 2024a).

Metode Empiris bekerja dengan memecah sebuah string menjadi beberapa kemungkinan kombinasi kata, kemudian setiap kombinasi tersebut dicocokkan dengan basis data kamus. Sebagai contoh, kata “makannasi” tidak ditemukan dalam kamus bahasa Indonesia karena ditulis tanpa spasi. Metode ini akan mencoba berbagai kemungkinan pemisahan kata hingga ditemukan kombinasi yang valid, seperti “makan” dan “nasi” yang keduanya terdaftar dalam basis data. Berdasarkan proses tersebut, sistem dapat memberikan saran perbaikan berupa “makan nasi”.

Proses pemecahan dilakukan dengan menambahkan spasi pada berbagai posisi yang memungkinkan sampai diperoleh hasil koreksi yang paling sesuai (Oktavia Suzanti et al., 2024a).

Berdasarkan latar belakang yang telah dijelaskan, penelitian ini bertujuan untuk mengidentifikasi dan mendeteksi kesalahan ejaan serta penulisan kata (*spelling error*) yang muncul pada teks informal Bahasa Indonesia, khususnya pada konten media sosial yang banyak mengandung bahasa tidak baku dan kata yang tidak sesuai ejaan. Penelitian ini menggunakan metode *Damerau Levenshtein Distance* untuk menghitung tingkat kemiripan antar kata dan mendeteksi kesalahan penulisan, serta metode Empiris untuk mengoptimalkan hasil koreksi kata berdasarkan pengamatan nyata terhadap bentuk penulisan yang umum digunakan oleh pengguna media sosial.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Pengoreksian ejaan secara otomatis menggunakan metode *Damerau Levenshtein Distance* memang mampu mendeteksi dan memperbaiki kesalahan pengetikan (*typo*) dan kesalahan ejaan pada teks. Namun, dalam konteks teks informal bahasa Indonesia, seperti komentar di media sosial atau pesan singkat, sering dijumpai bentuk penulisan yang tidak baku, seperti penulisan kata yang seharusnya terpisah menjadi satu yang tidak sesuai dengan kamus besar bahasa Indonesia (KBBI). Kondisi tersebut menyebabkan hasil koreksi dari metode *Damerau Levenshtein Distance* terkadang tidak akurat karena metode ini hanya mengukur jarak antar karakter tanpa mempertimbangkan konteks dan kebiasaan penulisan pengguna. Sehingga, diperlukan metode Empiris untuk menganalisis dan mengevaluasi hasil koreksi ejaan secara langsung terhadap data teks informal, sehingga diperoleh hasil yang lebih relevan, realistis, dan sesuai dengan pola penulisan pengguna sebenarnya.

1.2.2 Usulan Solusi

Solusi yang diusulkan terhadap permasalahan yang telah dijelaskan diatas, penelitian kali ini akan menerapkan metode *Damerau Levenshtein Distance* untuk mendeteksi dan memperbaiki kesalahan penulisan kata dan ejaan (*spelling error*) pada teks informal bahasa Indonesia dengan menghitung jarak kemiripan antar

karakter dari kata yang salah dengan kata acuan. Hasil dari perhitungan jarak tersebut akan digunakan untuk menentukan kata yang paling mendekati ejaan yang benar. Selanjutnya, metode *Empiris* menguji hasil koreksi secara langsung pada kumpulan data teks informal, seperti percakapan di media sosial atau pesan singkat, untuk menilai efektivitas metode yang digunakan.

1.2.3 Pertanyaan Penelitian

Seberapa efektif kombinasi algoritma *Damerau Levenshtein Distance* (DLD) dan metode *Empiris* dalam meningkatkan akurasi sistem *spelling correction* pada teks informal Bahasa Indonesia, khususnya dalam menangani kesalahan penulisan kata tanpa spasi?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Penelitian ini hanya berfokus pada identifikasi dan koreksi kesalahan ejaan (*spelling error*) pada teks informal bahasa Indonesia.
2. Jenis kesalahan yang ditangani hanya mencakup kesalahan pengetikan sederhana, seperti penambahan, penghapusan, atau penggantian huruf, tanpa memperhatikan konteks makna kalimat secara mendalam.
3. *Dataset* yang digunakan berupa teks informal bahasa Indonesia yang dikumpulkan dari sosial media, dan tidak mencakup bahasa daerah, bahasa asing, atau teks formal.
4. Data acuan yang digunakan Adalah kamus besar Bahasa Indonesia (KBBI).

1.4 Tujuan dan Manfaat

1.4.1 Tujuan Penelitian

1. Untuk mengetahui tingkat akurasi metode *Damerau Levenshtein Distance* dalam mendeteksi dan mengoreksi kesalahan ejaan (*spelling error*) pada teks informal bahasa Indonesia.
2. Untuk mengetahui pengaruh penerapan metode *Empiris* terhadap peningkatan akurasi hasil koreksi ejaan yang dihasilkan oleh metode *Damerau Levenshtein Distance*.

1.4.2 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis. Secara teoritis, penelitian ini dapat menjadi referensi dalam

pengembangan studi terkait *spelling correction* atau koreksi ejaan otomatis pada teks informal bahasa Indonesia menggunakan metode *Damerau Levenshtein Distance* dan Empiris. Secara praktis, hasil penelitian ini diharapkan dapat membantu pengembang sistem pemrosesan bahasa alami (NLP) dalam meningkatkan akurasi koreksi ejaan pada platform digital seperti media sosial atau aplikasi berbasis teks.

1.5 Sistematika Kepenulisan

Sistematika penulisan proposal untuk tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang, permasalahan, tujuan, manfaat, batasan penelitian, serta sistematika penulisan sebagai dasar dan ruang lingkup penelitian yang dilakukan.

BAB II KAJIAN PUSTAKA

Bab ini membahas teori dan konsep dasar yang mendukung penelitian, meliputi spelling error, preprocessing teks, spelling correction, algoritma Damerau Levenshtein Distance, metode Empiris, serta penelitian terdahulu yang relevan.

BAB III METODELOGI PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, meliputi arsitektur sistem, dataset, preprocessing, pengecekan dan filtering kata, penerapan algoritma Damerau Levenshtein Distance dan metode Empiris, serta evaluasi dan skenario pengujian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil pengujian dan pembahasan sistem berdasarkan metode penelitian yang telah diterapkan.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang diperoleh dari hasil penelitian serta saran yang dapat digunakan untuk pengembangan penelitian selanjutnya.