

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan layanan berbasis digital menghasilkan semakin banyak data teks melalui ulasan pengguna aplikasi PLN Mobile. Ulasan tersebut memuat pengalaman, kritik, keluhan, dan saran yang dapat dimanfaatkan sebagai sumber informasi untuk mengevaluasi kualitas layanan. Data ulasan PLN Mobile telah digunakan dalam penelitian analisis sentimen menggunakan *Lexicon VADER* dan *Naive Bayes* (Asri et al., 2022). Pendekatan leksikon bahasa Indonesia dan IndoBERT juga telah diterapkan pada ulasan PLN Mobile (Asri et al., 2025). *Spelling corrector* Peter Norvig juga pernah digunakan sebagai bagian dari *preprocessing* sebelum klasifikasi sentimen (Asri et al., 2025). Kondisi tersebut menunjukkan bahwa ulasan PLN Mobile berpotensi dikembangkan lebih lanjut untuk penanganan kualitas teks, termasuk kesalahan ejaan.

Data ulasan pengguna bersifat tidak terstruktur karena ditulis secara bebas dan tidak selalu mengikuti kaidah bahasa Indonesia baku. Kesalahan ketik, singkatan, *slang*, pengulangan huruf, dan variasi penulisan dapat menyebabkan kata dengan maksud sama direpresentasikan sebagai token yang berbeda. Penanganan kata salah tulis dapat memengaruhi kualitas pengolahan teks bahasa Indonesia (Setiabudi et al., 2021). Pengulangan karakter juga dapat menghasilkan variasi bentuk kata yang membutuhkan normalisasi (Mansur et al., 2024). Pada data penelitian ditemukan bentuk seperti “gk” yang merujuk pada “tidak”, “eror” yang merujuk pada “error”, serta “lemottt” sebagai variasi dari “lemot”. Kesalahan tipografi semacam ini dapat memengaruhi pengolahan teks otomatis (Lee et al., 2022).

Data penelitian terdiri atas 202.532 ulasan dengan total 2.398.573 token kata dan 46.435 kata unik. Hasil identifikasi awal memperoleh 271 pasangan kata *slang* atau kesalahan penulisan yang dipetakan menuju 148 kata formal unik. Data tersebut masih berupa teks mentah dan belum tersusun sebagai data yang secara jelas membedakan kata benar dan kata salah. Kondisi ini menunjukkan perlunya pembentukan *dataset*

spelling checker sebagai data terstruktur dan terlabel untuk proses pelatihan serta pengujian model. Pembentukan *dataset* dilakukan dengan mengidentifikasi kata dari ulasan PLN Mobile, menentukan pasangan kata salah dan kata benar, kemudian memberikan label pada masing-masing kata sehingga dapat digunakan sebagai data pelatihan dan pengujian model. Ketersediaan korpus kesalahan ejaan menjadi bagian penting dalam pengembangan sistem pemeriksa ejaan bahasa Indonesia (Yanfi et al., 2023). Pendekatan berbasis pasangan kata salah dan kata benar juga telah digunakan dalam pengembangan *spell checking* berbasis data (Toleu et al., 2022).

Pengembangan model setelah pembentukan *dataset spelling checker* memerlukan metode yang mampu mempelajari pola perbedaan antara kata benar dan kata salah. Pendekatan *deep learning* telah digunakan dalam berbagai tugas klasifikasi teks (Minaee et al., 2021). Penerapannya juga telah dilakukan dalam pengembangan *spell checker* bahasa Indonesia (Lubis et al., 2023). Model *Bidirectional LSTM* dengan *Multi-Head Attention* telah digunakan untuk mendeteksi kesalahan ejaan bahasa Indonesia (Yanfi et al., 2024). CNN menjadi salah satu metode yang relevan karena mampu mengenali pola lokal pada data teks (Soni et al., 2023). Penerapan CNN pada pemeriksaan ejaan bahasa Indonesia menghasilkan akurasi tertinggi sebesar 0,89 ketika seluruh tahapan *preprocessing* digunakan (Lubis et al., 2023).

Penelitian ini menggunakan *Character-Level Convolutional Neural Network* karena kesalahan ejaan banyak terjadi melalui perubahan karakter (Lubis et al., 2023; Yanfi et al., 2023). Kata “eror” dari “error” menunjukkan penghilangan karakter, sedangkan “lemott” dari “lemot” menunjukkan pengulangan karakter. Representasi pada tingkat karakter memungkinkan susunan huruf tetap dipertahankan sehingga perubahan lokal dapat dipelajari langsung oleh model (Al-Hussaeni et al., 2023). Pendekatan pada unit yang lebih kecil daripada kata juga telah digunakan untuk menangani kesalahan tipografi dan variasi karakter (Stankevičius et al., 2022). *Character-Level CNN* digunakan untuk mempelajari pola yang membedakan kata benar dan kata salah pada ulasan PLN Mobile.

Text preprocessing perlu dianalisis karena dapat memengaruhi representasi data yang diterima model. Tahapan tersebut dapat membantu membersihkan dan menyeragamkan teks, tetapi efektivitasnya bergantung pada karakteristik data dan

metode yang digunakan (Palomino et al., 2022). Perbedaan tahapan *preprocessing* diketahui dapat menghasilkan perubahan performa pada sistem *spell checker* (Lubis et al., 2023). Pada model berbasis karakter, proses normalisasi juga dapat menghilangkan informasi yang sebenarnya diperlukan untuk mendeteksi kesalahan. Normalisasi "lemottt" menjadi "lemot", misalnya, menghilangkan pola pengulangan karakter (Mansur et al., 2024). Penelitian ini menerapkan tahapan *text preprocessing*, baik tanpa maupun dengan *preprocessing* tambahan, dalam proses implementasi *spelling checker* berbasis *Character-Level CNN*.

Celah penelitian terlihat dari penelitian pada ulasan PLN Mobile yang masih lebih banyak diarahkan pada analisis sentimen. Analisis menggunakan *Lexicon VADER* dan *Naive Bayes* menghasilkan akurasi sebesar 70% (Asri et al., 2022). *Spelling corrector* Peter Norvig juga masih digunakan sebagai bagian dari *preprocessing* sebelum klasifikasi sentimen dan menghasilkan akurasi tertinggi sebesar 82% (Asri et al., 2025). Penelitian mengenai kesalahan ejaan bahasa Indonesia telah mengembangkan korpus kesalahan ejaan (Yanfi et al., 2023) dan model *deep learning* untuk deteksi kesalahan ejaan (Yanfi et al., 2024). Pembentukan *dataset spelling checker* berdasarkan karakteristik aktual ulasan PLN Mobile, penerapan *Character-Level CNN*, serta penerapan *text preprocessing* dalam proses implementasi *spelling checker* belum menjadi fokus utama pada penelitian terdahulu yang dirujuk.

Model yang dikembangkan perlu dievaluasi secara terukur agar kemampuannya dalam membedakan kata benar dan kata salah dapat diketahui secara objektif. Evaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score* telah diterapkan dalam penelitian *spell checker* bahasa Indonesia (Lubis et al., 2023). Keempat metrik tersebut digunakan untuk melihat ketepatan prediksi secara keseluruhan serta kemampuan model dalam mengenali masing-masing kelas. Hasil evaluasi juga digunakan untuk membandingkan performa model pada skenario dengan dan tanpa *preprocessing* tambahan. Urgensi penelitian terletak pada kebutuhan untuk menangani kesalahan ejaan pada ulasan PLN Mobile yang berjumlah besar dan bervariasi. Pemeriksaan manual terhadap 2.398.573 token kata tidak efisien sehingga diperlukan pembentukan *dataset spelling checker*, penerapan *Character-Level CNN*, serta evaluasi hasil model pada data tanpa *preprocessing* tambahan dan dengan *preprocessing* tambahan.

Penelitian ini diharapkan mendukung deteksi kesalahan ejaan pada ulasan PLN Mobile secara otomatis.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana membentuk *dataset spelling checker* Bahasa Indonesia dari ulasan PLN Mobile?
2. Bagaimana penerapan *Character-Level* CNN untuk mendeteksi kesalahan ejaan Bahasa Indonesia?
3. Bagaimana penerapan *text preprocessing* pada proses implementasi *spelling checker*?
4. Bagaimana hasil evaluasi model berdasarkan *Accuracy*, *Precision*, *Recall* dan *F1-Score*?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, tujuan dari penelitian ini adalah sebagai berikut:

1. Membentuk *dataset* kesalahan ejaan Bahasa Indonesia dari ulasan PLN Mobile.
2. Mengimplementasikan model *Character-Level* CNN untuk mendeteksi kata benar dan kata salah.
3. Menerapkan tahapan *text preprocessing* dalam proses implementasi *spelling checker* Bahasa Indonesia.
4. Mengevaluasi model berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat, baik secara teoretis, praktis, maupun bagi pengembangan sistem.

1. Manfaat Teoretis

Penelitian ini diharapkan dapat memberikan kontribusi dalam bidang *Natural Language Processing* (NLP) Bahasa Indonesia, khususnya pada pengembangan *spelling checker* berbasis *text preprocessing* dan model *deep*

learning. Selain itu, penelitian ini juga dapat menjadi referensi bagi penelitian selanjutnya yang membahas deteksi kesalahan ejaan, normalisasi kata, serta pengolahan teks informal berbahasa Indonesia.

2. Manfaat Praktis

Penelitian ini diharapkan dapat membantu meningkatkan kualitas data teks ulasan pengguna aplikasi PLN *Mobile* melalui proses deteksi kesalahan ejaan. Dengan adanya *spelling checker*, kata yang salah tulis, tidak baku, atau mengandung variasi penulisan informal dapat dikenali sehingga data teks menjadi lebih rapi dan lebih siap digunakan untuk proses pengolahan bahasa selanjutnya.

3. Manfaat bagi Pengembangan Sistem

Metode yang digunakan dalam penelitian ini dapat dijadikan sebagai acuan dalam pengembangan sistem *spelling checker* Bahasa Indonesia berbasis *deep learning*. Sistem tersebut dapat dikembangkan lebih lanjut untuk mendukung proses normalisasi teks, klasifikasi teks, analisis opini, maupun aplikasi NLP lainnya yang membutuhkan data teks yang bersih dan terstruktur.

1.5 Batasan Masalah

Guna menjaga fokus penelitian agar tujuan dapat tercapai secara spesifik dan terukur, ruang lingkup penelitian ini dibatasi pada aspek-aspek berikut:

1. Data yang digunakan berasal dari ulasan pengguna aplikasi PLN *Mobile*.
2. *Dataset* yang digunakan mengacu pada data ulasan PLN *Mobile*, yaitu sekitar 437.683 entri ulasan awal dan 202.532 ulasan bersih setelah proses pembersihan data.
3. Penelitian ini berfokus pada *spelling checker* Bahasa Indonesia, bukan pada analisis sentimen.
4. Kesalahan yang dibahas dalam penelitian ini meliputi kesalahan penulisan kata, kata tidak baku, slang, singkatan, dan variasi penulisan informal yang ditemukan pada data ulasan pengguna.
5. Tahapan *preprocessing* dalam penelitian ini bertujuan untuk membersihkan dan menyeragamkan data awal tanpa menghilangkan informasi kesalahan ejaan. Kata yang mengandung *Typo*, slang, singkatan, dan variasi penulisan

informal tetap dipertahankan pada pembentukan *dataset spelling checker* agar model CNN dapat mempelajari pola kesalahan karakter.

6. *Dataset spelling checker* yang dibentuk terdiri dari label kata benar dan kata salah.
7. Model *deep learning* yang digunakan dalam penelitian ini adalah *Convolutional Neural Network* (CNN) sebagai model utama untuk klasifikasi kata benar dan kata salah.
8. Evaluasi performa model dilakukan menggunakan metrik akurasi, presisi, *recall*, dan F1-score.
9. Penelitian ini tidak membahas klasifikasi sentimen positif, negatif, dan netral.
10. Penelitian ini tidak berfokus pada perbaikan tata bahasa atau struktur kalimat, tetapi hanya pada deteksi kesalahan ejaan pada tingkat kata.