Pendekatan Rule-based Menggunakan Kamus dan Named Entity Recognizer untuk Mendeteksi dan Mengoreksi Kesalahan Penulisan Huruf Kapital pada Teks Berbahasa Indonesia = A Rule-based Approach Using Dictionary and Named Entity Recognizer for Detecting and Correcting Capitalization Errors in Indonesian Text

Sultan Daffa Nusantara, author

Abstrak

Penggunaan huruf kapital merupakan aspek penting dalam menulis bahasa Indonesia yang baik dan benar. Aturan penggunaan huruf kapital dalam bahasa Indonesia telah dijelaskan dalam Pedoman Umum Ejaan Bahasa Indonesia (PUEBI) yang terdiri dari 23 aturan. Penelitian sebelumnya telah memulai mengembangkan pendeteksi dan pengoreksi kesalahan huruf kapital untuk bahasa Indonesia menggunakan pendekatan rule-based dengan kamus dan komponen Named Entity Recognition (NER). Namun, penelitian tersebut hanya mencakup 9 dari 23 aturan huruf kapital yang tercantum dalam PUEBI dan dataset uji yang digunakan tidak dipublikasikan sehingga tidak dapat digunakan untuk penelitian selanjutnya. Penelitian ini bertujuan untuk mengusulkan metode untuk mendeteksi dan mengoreksi 14 dari 23 aturan PUEBI menggunakan pendekatan yang mirip dengan penelitian sebelumnya. Model NER dikembangkan menggunakan pretrained language model IndoBERT yang dilakukan fine-tuning dengan dataset NER. Untuk menguji metode rule-based yang diusulkan, dibuat sebuah dataset sintesis yang terdiri dari 5.000 pasang kalimat. Setiap pasang terdiri dari kalimat benar secara aturan huruf kapital dan padanan kalimat salahnya. Kalimat salah dibuat dengan mengubah beberapa huruf kapital di kalimat yang awalnya benar. Sebelum dilakukan perbaikan terhadap kalimat yang salah, didapatkan akurasi sebesar 83,10%. Namun, setelah menggunakan metode ini, tingkat akurasi meningkat 12,35% menjadi 95,45%.
.....The correct use of capital letters plays a vital role in writing well-formed and accurate Indonesian sentences. Pedoman Umum Ejaan Bahasa Indonesia (PUEBI) provide a comprehensive set of 23 rules that explain how to use capital letters correctly. Previous research has attempted to develop a rule-based system to detect and correct capital letter errors in Indonesian text using dictionaries and Named Entity Recognition (NER). However, this study only covered 9 out of the 23 capital letter rules specified in PUEBI, and the test dataset used was not publicly available for further analysis. In this study, we aim to propose a method that can identify and rectify 14 out of the 23 PUEBI rules, following a similar approach to previous research. The NER model was trained using the IndoBERT pretrained language model and fine-tuned with a specific NER dataset. To evaluate the effectiveness of our rule-based method, we created a synthetic dataset comprising 5,000 sentence pairs. Each pair consists of a correctly capitalized sentence and an equivalent sentence with incorrect capitalization. Before applying our method, the baseline accuracy was 83.10%. However, after implementing our approach, the accuracy improved by 12.35% to reach 95.45%.