

BOILERPLATE REMOVAL UNTUK KONTEN UTAMA JURNAL ILMIAH MENGGUNAKAN BOILERNET

Nama Mahasiswa : Felix Dean Janitra
NIM : 11211040
Dosen Pembimbing Utama : Gusti Ahmad Fanshuri Alfarisy, S.Kom.,
M.Kom.
Dosen Pembimbing Pendamping : Aninditya Anggari Nuryono, S.T., M.Eng.

ABSTRAK

Di era digital, jumlah publikasi jurnal ilmiah yang tersedia secara daring semakin meningkat pesat. Namun, proses ekstraksi informasi dari jurnal ilmiah sering terganggu oleh elemen *boilerplate*, seperti header, footer, navigasi, dan metadata yang tidak berkaitan langsung dengan konten utama. Keberadaan elemen-elemen ini menghambat analisis teks dan pemrosesan lebih lanjut dalam berbagai aplikasi akademik. Penelitian ini bertujuan untuk mengembangkan metode penghapusan boilerplate secara otomatis menggunakan *BoilerNet*, sebuah model berbasis deep learning dengan pendekatan *Bi-LSTM*. Model ini dikembangkan untuk mengklasifikasikan setiap bagian teks dalam halaman jurnal dan memisahkan konten utama dari *boilerplate* secara lebih akurat. Metode yang digunakan dalam penelitian ini melibatkan beberapa tahap, termasuk pengumpulan dataset jurnal ilmiah melalui *web scraping*, *preprocessing* data untuk normalisasi dan pelabelan, serta pelatihan model dengan optimasi menggunakan *Binary Cross Entropy*. Evaluasi model dilakukan menggunakan metrik seperti *Precision*, *Recall* dan *F1-score* untuk mengukur efektivitas model dalam menghapus elemen boilerplate. Pengujian dilakukan menggunakan total 180 data, terdistribusi menjadi 50 data pelatihan, 30 validasi, dan 100 pengujian. Hasil eksperimen menunjukkan bahwa konfigurasi optimal *BoilerNet* dengan 256 *hidden unit*, *dropout rate* 0.4, dua lapisan *LSTM*, dan ukuran *dense layer* 256 mencatat *F1 score* 0.91 untuk kelas *boilerplate* (negatif) dan 0.81 untuk kelas konten utama (positif). *Average F1 score* keseluruhan mencapai 0.863. Hasil ini mengindikasikan bahwa *BoilerNet* mampu meningkatkan performa ekstraksi konten utama secara signifikan dibandingkan metode konvensional seperti *ReadabilityJS*.

Kata kunci :

Boilerplate Removal, *BoilerNet*, *Deep Learning*, *Bi-LSTM*, Jurnal Ilmiah, Ekstraksi Konten Utama

www.itk.ac.id



(Halaman ini sengaja dikosongkan)

www.itk.ac.id