ABSTRACT

ANALYSIS OF OPTIMAL PARAMETERS IN DETECTING I-DIS LABEL ERRORS IN NAMED ENTITY RECOGNITION OF INFECTIOUS DISEASE NEWS HEADLINES BASED ON A HYBRID INDOBERT–BILSTM–CRF MODEL

By

Anita Caroline Gunawan

Entity extraction in health news texts is an important component in supporting the automated analysis of infectious disease information. However, the characteristics of news headlines, which are concise and context-dense, pose challenges for entity labeling using the Named Entity Recognition (NER) approach. This study aims to develop an NER model based on a hybrid IndoBERT–BiLSTM–CRF architecture by integrating Part-of-Speech (POS) tagging features using Stanza and FastText embeddings to recognize disease entities in Indonesian news headlines, determine the optimal hyperparameter configuration, and analyze labeling errors, particularly in the I-DIS label. The research method includes data preprocessing, annotation using the BIO scheme, model training with several hyperparameter combinations, and evaluation using accuracy, precision, recall, F1-score, confusion matrix, and ROC–AUC metrics. The results show that the best-performing model achieves an accuracy of 98.26% and a weighted F1-score of 98.28%, and is capable of recognizing disease entities with an accuracy of approximately 99%. The optimal configuration is obtained with a hidden dimension of 256, one BiLSTM layer, a dropout rate of 0.22, a learning rate of 2.53×10^{-5}, and a batch size of 16. Error analysis indicates that the primary weakness lies in the I-DIS label due to data imbalance; however, it does not significantly affect the overall model performance in detecting disease entities. Overall, the hybrid IndoBERT–BiLSTM–CRF model, supported by POS tagging (Stanza) and FastText embeddings, proves to be effective in recognizing disease entities in Indonesian news headlines and demonstrates competitive performance in NER tasks within the healthcare domain.

Keywords: Named Entity Recognition, IndoBERT–BiLSTM-CRF, POS Tagging, FastText Embedding, Error Analysis.

ABSTRAK

ANALISIS PARAMETER TERBAIK DALAM DETEKSI KESALAHAN LABEL I-DIS PADA *NAMED ENTITY RECOGNITION* JUDUL BERITA PENYAKIT MENULAR BERBASIS *HYBRID* INDOBERT–BILSTM–CRF

Oleh

Anita Caroline Gunawan

Ekstraksi entitas pada teks berita kesehatan merupakan komponen penting dalam mendukung analisis otomatis informasi penyakit menular. Namun, karakteristik judul berita yang singkat dan padat konteks menimbulkan tantangan dalam pelabelan entitas menggunakan pendekatan *Named Entity Recognition* (NER). Penelitian ini bertujuan membangun model NER berbasis arsitektur hibrida IndoBERT–BiLSTM–CRF dengan integrasi fitur POS *tagging* menggunakan Stanza dan *embedding* FastText untuk mengenali entitas penyakit pada judul berita berbahasa Indonesia, menentukan konfigurasi *hyperparameter* terbaik, serta menganalisis kesalahan pelabelan, khususnya pada label I-DIS. Metode penelitian meliputi praproses data, anotasi dengan skema BIO, pelatihan model menggunakan beberapa kombinasi *hyperparameter*, dan evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, F1-*score*, *confusion matrix*, dan ROC–AUC. Hasil penelitian menunjukkan bahwa model terbaik mencapai *accuracy* sebesar 98,26% dan *weighted* F1-*score* sebesar 98,28%, serta mampu mengenali entitas penyakit dengan tingkat ketepatan sekitar 99%. Konfigurasi optimal diperoleh pada *hidden dimension* 256, satu lapisan BiLSTM, *dropout* 0,22, *learning rate* $2{,}53 \times 10^{-5}$, dan *batch size* 16. Analisis kesalahan menunjukkan bahwa kelemahan utama terdapat pada label I-DIS akibat ketidakseimbangan distribusi data, namun tidak memberikan dampak signifikan terhadap kinerja model dalam mendeteksi entitas penyakit secara keseluruhan. Secara keseluruhan, model hibrida IndoBERT–BiLSTM–CRF dengan dukungan fitur POS *tagging* (Stanza) serta *embedding* FastText terbukti efektif dalam mengenali entitas penyakit pada judul berita berbahasa Indonesia dan menunjukkan kinerja yang kompetitif pada tugas NER di domain kesehatan.

Kata-kata kunci: *Named Entity Recognition*, IndoBERT–BiLSTM-CRF, POS *Tagging*, FastText *Embedding*, *Error Analysis*.