ABSTRACT

ANALYSIS OF THE EFFECT OF LATENT DIRICHLET ALLOCATION AND GRAPH CONVOLUTIONAL NETWORK ON MODEL X FOR NAMED ENTITY RECOGNITION OF INDONESIAN-LANGUAGE INFECTIOUS DISEASE NEWS

By

Erin Elfitriani

The increasing volume of Indonesian digital text has driven the need for automatic information extraction systems, particularly Named Entity Recognition (NER). Class imbalance and the complexity of entity structures remain challenges that may affect the consistency of entity type recognition. This study aims to analyze the performance of Model X, namely the IndoBERT–BiLSTM hybrid model, and to evaluate the effect of integrating Latent Dirichlet Allocation (LDA) topic features and the structural relations of Graph Convolutional Network (GCN) on NER performance based on entity types. The evaluation was conducted using precision, recall, F1-score, accuracy, and macro-average F1-score metrics. The results show that Model X achieved a macro F1-score of 0.9127 with an accuracy of 0.9681. The integration of LDA improved the recall value but reduced precision, resulting in a macro F1-score of 0.7119. The model integrated with GCN demonstrated a better balance between precision and recall, achieving a macro F1-score of 0.8611. Meanwhile, the combination of LDA and GCN produced high recall across all entity types, but the decline in precision led to a macro F1-score of 0.6709. These findings indicate differences in performance characteristics across the model scenarios, where Model X yielded the highest aggregate performance, while the integration of GCN showed more consistent entity detection capability compared with the other approaches.

Keywords: Named Entity Recognition, IndoBERT, BiLSTM, Latent Dirichlet Allocation, Graph Convolutional Network

ABSTRAK

ANALISIS PENGARUH *LATENT DIRICHLET ALLOCATION* DAN *GRAPH CONVOLUTIONAL NETWORK* PADA MODEL X UNTUK *NAMED ENTITY RECOGNITION* BERITA PENYAKIT MENULAR BERBAHASA INDONESIA

Oleh

Erin Elfitriani

Peningkatan volume teks digital berbahasa Indonesia mendorong kebutuhan akan sistem ekstraksi informasi otomatis, khususnya *Named Entity Recognition* (NER). Ketidakseimbangan kelas dan kompleksitas struktur entitas masih menjadi tantangan yang dapat memengaruhi konsistensi pengenalan tipe entitas. Penelitian ini bertujuan untuk menganalisis kinerja model X, yaitu model hibrida IndoBERT–BiLSTM, serta mengevaluasi pengaruh integrasi fitur topik *Latent Dirichlet Allocation* (LDA) dan relasi struktural *Graph Convolutional Network* (GCN) terhadap performa NER berdasarkan tipe entitas. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, *accuracy*, dan *macro average F1-score*. Hasil penelitian menunjukkan bahwa model X memperoleh *macro F1-score* sebesar 0,9127 dengan akurasi 0,9681. Integrasi LDA meningkatkan nilai *recall*, namun menurunkan *precision* sehingga *macro F1-score* menjadi 0,7119. Model dengan integrasi GCN menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall* dengan *macro F1-score* sebesar 0,8611. Sementara itu, kombinasi LDA dan GCN menghasilkan *recall* yang tinggi pada seluruh tipe entitas, tetapi penurunan *precision* menyebabkan *macro F1-score* menjadi 0,6709. Hasil penelitian ini menunjukkan adanya perbedaan karakteristik kinerja antar skenario model, di mana model X memberikan performa agregat tertinggi, sedangkan integrasi GCN menunjukkan kemampuan deteksi entitas yang lebih konsisten dibandingkan pendekatan lainnya.

Kata-kata kunci: *Named Entity Recognition*, IndoBERT, BiLSTM, *Latent Dirichlet Allocation*, *Graph Convolutional Network*