ABSTRAK

Pertumbuhan publikasi ilmiah menuntut adanya sistem klasifikasi topik otomatis yang presisi. Penelitian ini mengusulkan model *deep learning* hibrida untuk mengklasifikasikan bidang studi (*Field of Study*) makalah ilmiah. Model *pre-trained* SciBERT digunakan untuk mengekstraksi vektor fitur semantik dari teks judul dan abstrak menjadi matriks *embedding* 768 dimensi. Selanjutnya, algoritma GraphSAGE diimplementasikan untuk melakukan agregasi fitur spasial dengan memanfaatkan struktur topologi jaringan sitasi (*citation network*). Eksperimen dilakukan pada dataset DBLP Citation Network dengan rentang publikasi 1990-2020, di mana data dibagi menjadi 80% data latih, 10% data validasi, dan 10% data uji. Pengolahan komputasi data berskala besar dioptimalkan menggunakan teknik pengambilan sampel tetangga (*neighbor sampling*) dan penarikan fitur HDF5 dinamis (*Dynamic HDF5 Fetching*) pada setiap proses *mini-batch*. Evaluasi final model pada himpunan data uji (*test set*) menunjukkan kinerja klasifikasi pada Top 20 FOS mencapai *Accuracy* 62.17%, *Precision* 0.5392, *Recall* 0.4757, dan *F1-Score* 0.4853. Pada skenario kelas Top 10 FOS, model memperoleh *Accuracy* 72.54%, *Precision* 0.6835, *Recall* 0.6310, dan *F1-Score* 0.6495. Sedangkan pada penyempitan kelas Top 5 FOS, model berhasil mencapai performa tertinggi dengan *Accuracy* 82.83%, *Precision* 0.8184, *Recall* 0.8131, dan *F1-Score* 0.8144. Penggabungan antara representasi semantik SciBERT dan agregasi graf GraphSAGE terbukti efektif menghasilkan sistem klasifikasi multikelas yang stabil dan akurat.

Kata Kunci : Klasifikasi Topik, Artikel Ilmiah, *SciBERT, GraphSAGE, Citation Network.*

ABSTRACT

The growth of scientific publications requires a precise automatic topic classification system. This research proposes a hybrid deep learning model to classify the Field of Study of scientific papers. The pre-trained SciBERT model is utilized to extract semantic feature vectors from title and abstract texts into a 768-dimensional embedding matrix. Furthermore, the GraphSAGE algorithm is implemented to perform spatial feature aggregation by leveraging the topological structure of the citation network. Experiments were conducted on the DBLP Citation Network dataset spanning from 1990 to 2020, where the data was randomly split into 80% training data, 10% validation data, and 10% testing data. Large-scale data computation is optimized using neighbor sampling techniques and Dynamic HDF5 Fetching during every mini-batch process. Final model evaluation on the test set demonstrates that classification performance on the Top 20 FOS reaches an Accuracy of 62.17%, Precision of 0.5392, Recall of 0.4757, and F1-Score of 0.4853. In the Top 10 FOS class scenario, the model obtained an Accuracy of 72.54%, Precision of 0.6835, Recall of 0.6310, and F1-Score of 0.6495. Meanwhile, in the Top 5 FOS class, the model achieved its peak performance with an Accuracy of 82.83%, Precision of 0.8184, Recall of 0.8131, and F1-Score of 0.8144. The combination of SciBERT's semantic representation and GraphSAGE's graph aggregation has proven effective in producing a stable and accurate multiclass classification system.

***Keywords**: Topic Classification, Scientific Articles, SciBERT, GraphSAGE, Citation Network.*