Abstrak

Dalam NLP, kata kunci menjadi krusial untuk melakukan information retrieval (IR) dan document summarization. Kendati demikian, identifikasi kata kunci dari dokumen besar menjadi sangat sulit jika dilakukan secara manual. Oleh karena itu, ekstraksi kata kunci otomatis menjadi penting untuk mengatasi volume informasi yang meningkat, terutama di domain akademis. Metode ekstraksi kata kunci saat ini cenderung menggunakan fitur semantik global atau statistik lokal secara terpisah, menghasilkan performa rendah. Oleh karena itu, diperlukan pendekatan yang menggabungkan informasi statistik lokal dengan model embedding untuk mencapai identifikasi kata kunci yang lebih baik. Dalam konteks akademis, hal ini dapat menyederhanakan proses kategorisasi publikasi ilmiah dan meningkatkan sistem IR di repositori makalah. Paper ini mengusulkan metode yang mengkombinasikan fitur statistik lokal dengan model embedding untuk ekstraksi kata kunci dari publikasi ilmiah. Pendekatan ini, diuji menggunakan model SciBERT dan klasifikasi SVM, berhasil melampaui metode sebelumnya dengan F-score sebesar 0.70 pada dataset SemEval2017. Penggabungan informasi statistik lokal dan semantik kontekstual ternyata krusial dalam mengidentifikasi kata kunci. Penelitian ini menegaskan bahwa kombinasi informasi statistik lokal dan informasi semantik kontekstual memiliki peran penting dalam identifikasi kata kunci.