ABSTRACT

Text analysis methods using basic word embedding models offer better performance than the bag-of-words-based methods. Unfortunately, the model yields poor performance for solving domain-specific NLP problems. Meanwhile, contextual word embedding has the advantage of dealing with word ambiguity which requires expensive computational resources and needs to be tested on specific tasks.

In this study, we conduct an experiment to combine word embedding and clustering method on science mapping using the abstract of scientific papers dataset. The word embedding used in this study is BERT, while the clustering algorithm is Fuzzy C-Means. BERT was fine-tuned on the text classification task and tested against the text clustering performance. In this study, we also experiment to examine the effect of contextual word embedding on text clustering performance.

Based on the experiment, fine-tuning BERT in domain-specific classification tasks can improve the text clustering performance. Meanwhile, the characteristics of BERT as contextual word embedding are not proven to have a positive effect on text clustering performance. Based on the results of the clustering evaluation with a target of 18 clusters, BERTBASE+DS_SKRIPSI-18 model got the highest score of F1 = 0.478, JC = 0.314, FM = 0.508, and RI = 0.888, using the EMBCLS method.

Keywords : word embedding, clustering, Fuzzy C-Means, TF-IDF, cluster labeling, BERT, transformers, science mapping

INTISARI

Metode analisis teks menggunakan model dasar *word embedding* memberikan performa yang lebih baik daripada metode berbasis *bag-of-words* akan tetapi kurang optimal untuk menyelesaikan permasalahan NLP di bidang pengetahuan tertentu. Sementara itu, *contextual word embedding* yang memiliki keunggulan dalam menangani ambiguitas kata membutuhkan sumber daya komputasi yang mahal sehingga keunggulan tersebut perlu diuji pada tugas tertentu.

Pada penelitian ini, dilakukan percobaan menggabungkan *word embedding* dengan algoritme *clustering* untuk melakukan pemetaan bidang ilmu menggunakan dataset abstrak penelitian yang didapatkan dari ETD UGM. *Word embedding* yang digunakan pada penelitian ini adalah BERT, sedangkan algoritme *clustering* yang digunakan adalah Fuzzy C-Means. Pada BERT diberikan tambahan pengetahuan untuk meningkatkan kemampuan klasifikasi teks pada 18 bidang ilmu dengan metode *fine-tuning* dan diuji pengaruhnya terhadap performa *clustering* abstrak penelitian. Pada penelitian ini juga dilakukan percobaan untuk menguji pengaruh *contextual word embedding* terhadap performa *clustering* abstrak penelitian.

Berdasarkan hasil percobaan yang dilakukan, BERT yang dilatih dengan kemampuan klasifikasi teks pada bidang ilmu tertentu terbukti mampu meningkatkan performa *clustering* abstrak penelitian. Sementara itu, karakteristik *contextual word embedding* dari BERT tidak terbukti memberikan pengaruh positif pada performa *clustering* abstrak penelitian. Berdasarkan hasil evaluasi *clustering* dengan target jumlah kelompok sebanyak 18 kelompok, $BERT_{BASE}$+DS_SKRIPSI-18 memperoleh nilai performa tertinggi, yaitu: skor F1=0.478, JC=0.314, FM=0.508, dan RI=0.888 menggunakan metode EMBCLS.

Kata kunci – word embedding, clustering, Fuzzy C-Means, TF-IDF, cluster labeling, BERT, transformers, science mapping