Analisis Kinerja Gabungan Metode Representasi Teks BERT dan Metode Clustering DEC untuk Pendeteksian Topik = Performance Analysis of BERT as Text Representation Method and DEC Clustering Method for Topic Detection

Lista Kurniawati, author

Abstrak

Pendeteksian topik merupakan masalah komputasi yang menganalisis kata-kata dari suatu data teks untuk menemukan topik yang ada di dalam teks tersebut. Pada data yang besar, pendeteksian topik lebih efektif dan efisien dilakukan dengan metode machine learning. Data teks harus diubah ke dalam bentuk representasi vektor numeriknya sebelum dimasukkan ke model machine learning. Metode representasi teks yang umum digunakan adalah TF-IDF. Namun, metode ini menghasilkan representasi data teks yang tidak memperhatikan konteksnya. BERT (Bidirectional Encoder Representation from Transformer) merupakan metode representasi teks yang memperhatikan konteks dari suatu kata dalam dokumen. Penelitian ini membandingkan kinerja model BERT dengan model TF-IDF dalam melakukan pendeteksian topik. Representasi data teks yang diperoleh kemudian dimasukkan ke model machine learning. Salah satu metode machine learning yang dapat digunakan untuk menyelesaikan masalah pendeteksian topik adalah clustering. Metode clustering yang populer digunakan adalah Fuzzy C-Means. Namun, metode Fuzzy C-Means tidak efektif pada data berdimensi tinggi. Karena data teks berita biasanya memiliki ukuran dimensi yang cukup tinggi, maka perlu dilakukan proses reduksi dimensi. Saat ini, terdapat metode clustering yang melakukan reduksi dimensi berbasis deep learning, yaitu Deep Embedded Clustering (DEC). Pada penelitan ini digunakan model DEC untuk melakukan pendeteksian topik. Eksperimen pendeteksian topik menggunakan model DEC (member) dengan metode representasi teks BERT pada data teks berita menunjukkan nilai coherence yang sedikit lebih baik dibandingkan dengan menggunakan metode representasi teks TF-IDF.
.....
Topic detection is a computational problem that analyzes words of a textual data to find the topics in it. In large data, topic detection is more effective and efficient using machine learning methods. Textual data must be converted into its numerical vector representation before being entered into a machine learning model. The commonly used text representation method is TF-IDF. However, this method produces a representation of text data that does not consider the context. BERT (Bidirectional Encoder Representation from Transformers) is a text representation method that pays attention to the context of a word in a document. This study compares the performance of the BERT model with the TF-IDF model in detecting topics. The representation of the text data obtained is then entered into the machine learning model. One of the machine learning methods that can be used to solve topic detection problems is clustering. The popular clustering method used is Fuzzy CMeans. However, the Fuzzy C-Means method is not effective on high-dimensional data. Because news text data usually has a high dimension, it is necessary to carry out a dimension reduction process. Currently, there is a clustering method that performs deep learning-based dimension reduction, namely Deep Embedded Clustering (DEC). In this research, the DEC model is used to detect topics. The topic detection experiment using the DEC (member) model with the BERT text representation method on news text data shows a slightly better coherence value than using the TF-IDF text representation method.