Analisis Performa EFCM dengan BERT sebagai Representasi Teks pada Pendeteksian Topik = The Performance of EFCM with BERT as Text Representation on Topic Detection

Nicholas Ramos Richardo, author

Abstrak

Pendeteksian topik adalah suatu proses untuk menentukan suatu topik dalam teks dengan menganalisis kata di dalam teks tersebut. Pendeteksian topik dapat dilakukan dengan membaca isi dari teks tersebut. Namun, cara ini semakin sulit apabila data yang dimiliki semakin besar. Memanfaatkan metode machine learning dapat menjadi alternatif dalam menangani data yang berjumlah besar. Metode clustering adalah metode pengelompokkan data yang mirip dari suatu kumpulan data. Beberapa contoh metode clustering adalah K-Means, Fuzzy C-Means (FCM), dan Eigenspaced-Based Fuzzy C-Means (EFCM). EFCM adalah metode clustering yang memanfaatkan metode reduksi dimensi Truncated Singular Value Decomposition (TSVD) dengan metode FCM (Murfi, 2018). Dalam pendeteksian topik, teks harus direpresentasikan kedalam bentuk vektor numerik karena model clustering tidak dapat memproses data yang berbetuk teks. Metode yang sebelumnya umum digunakan adalah Term-Frequency Inversed Document Frequency (TFIDF). Pada tahun 2018 diperkenalkan suatu metode baru yaitu metode Bidirectional Encoder Representations from Transformers (BERT). BERT merupakan pretrained language model yang dikembangkan oleh Google. Penelitian ini akan menggunakan model BERT dan metode clutering EFCM untuk masalah pendeteksian topik. Kinerja performa model dievaluasi dengan menggunakan metrik evaluasi coherence. Hasil simulasi menunjukkan penentuan topik dengan metode modifikasi TFIDF lebih unggul dibandingkan dengan metode centroid-based dengan dua dari tiga dataset yang digunakan metode modifikasi TFIDF memiliki nilai coherence yang lebih besar. Selain itu, BERT lebih unggul dibandingkan dengan metode TFIDF dengan nilai coherence BERT pada ketiga dataset lebih besar dibandingkan dengan nilai coherence TFIDF.
.....Topic detection is a process to determine a topic in the text by analyzing the words in the text. Topic detection can be done with reading the contents of the text.However, this method is more difficult when bigger data is implemented. Utilizing machine learning methods can be an alternative approach for handling a large amount of data. The clustering method is a method for grouping similar data from a data set. Some examples of clustering methods are K-Means, Fuzzy C-Means (FCM), and Eigenspaced-Based Fuzzy C-Means (EFCM). EFCM is a clustering method that utilizes the truncated dimension reduction method Singular Value Decomposition (TSVD) with the FCM method (Murfi, 2018). In topic detection, the text must be represented in numerical vector form because the clustering model cannot process data in the form of text. The previous method that was most commonly used is the Term-Frequency Inverse Document Frequency (TFIDF). In 2018 a new method was introduced, namely the Bidirectional Encoder method Representations from Transformers (BERT). BERT is a pretrained language model developed by Google. This study will use the BERT model and the EFCM clustering method for topic detection problems. The performance of the model is evaluated using the coherence evaluation metric. The simulation results show that modified TFIDF method for topic determination is superior to the centroid-based method with two of the three datasets used by modified TFIDF method having a greater coherence value. In addition, BERT is superior to the TFIDF method with the BERT coherence value in the three datasets greater than the TFIDF

coherence value.