Analisis kinerja BERT sebagai metode representasi teks untuk text clustering = Performance analysis of BERT as a text representation method for text clustering.

Alvin Subakti, author

Abstrak

Text clustering adalah teknik pengelompokan teks sehingga teks di dalam kelompok yang sama memiliki tingkat similaritas yang lebih tinggi satu sama lain dibandingkan dengan teks pada kelompok yang berbeda. Proses pengelompokkan teks secara manual membutuhkan waktu dan sumber daya yang banyak sehingga digunakan machine learning untuk melakukan pengelompokan secara otomatis. Representasi dari teks perlu diekstraksi sebelum dimasukkan ke dalam model machine learning. Metode yang umumnya digunakan untuk mengekstraksi representasi data teks adalah TFIDF. Namun, metode TFIDF memiliki kekurangan yaitu tidak memperhatikan posisi dan konteks penggunaan kata. Model BERT adalah model yang dapat menghasilkan representasi kata yang bergantung pada posisi dan konteks penggunaan suatu kata dalam kalimat. Penelitian ini menganalisis kinerja model BERT sebagai metode representasi data teks dengan membandingkan model BERT dengan TFIDF. Selain itu, penelitian ini juga mengimplementasikan dan membandingkan kinerja metode ekstraksi dan normalisasi fitur yang berbeda pada representasi teks yang dihasilkan model BERT. Metode ekstraksi fitur yang digunakan adalah max dan mean pooling. Sementara itu, metode normalisasi fitur yang digunakan adalah identity, layer, standard, dan min-max normalization. Representasi teks yang diperoleh dimasukkan ke dalam 4 algoritma clustering berbeda, yaitu k-means clustering, eigenspace-based fuzzy c-means, deep embedded clustering, dan improved deep embedded clustering. Kinerja representasi teks dievaluasi dengan menggunakan metrik clustering accuracy, normalized mutual information, dan adjusted rand index. Hasil simulasi menunjukkan representasi data teks yang dihasilkan model BERT mampu mengungguli representasi yang dihasilkan TFIDF pada 28 dari 36 metrik. Selain itu, implementasi ekstraksi dan normalisasi fitur yang berbeda pada model BERT memberikan kinerja yang berbeda-beda dan perlu disesuaikan dengan algoritma yang digunakan.
.....Text clustering is a task of grouping a set of texts in a way such that text in the same group will be more similar toward each other than to those from different group. The process of grouping text manually requires significant amount of time and labor. Therefore, automation utilizing machine learning is necessary. Text representation needs to be extracted to become the input for machine learning models. The common method used to represent textual data is TFIDF. However, TFIDF cannot consider the position and context of a word in a sentence. BERT model has the capability to produce text representation that incorporate position and context of a word in a sentence. This research analyzed the performance of BERT model as a text representation method by comparing it with TFIDF. Moreover, various feature extraction and normalization methods are also applied in text representation from BERT model. Feature extraction methods used are max and mean pooling. On the other hand, feature normalization methods used are identity, layer, standard, and min-max normalization. Text representation obtained become an input for 4 clustering algorithms, k-means clustering, eigenspace-based fuzzy c-means, deep embedded clustering, and improved deep embedded clustering. Performance of text representations in text clustering are evaluated utilizing clustering accuracy, normalized mutual information, and adjusted rand index. Simulation results showed that text representation

obtained from BERT model outperforms representation from TFIDF in 28 out of 36 metrics. Furthermore, different feature extraction and normalization produced varied performances. The usage of these feature extraction and normalization must be altered depending on the text clustering algorithm used.