Analisis Kinerja Metode BERT-IDEC untuk Deteksi Topik = BERT-IDEC Method Performance Analysis for Topic Detection

Syach Riyan Muhammad Ardiyansyah, author

Abstrak

Pendeteksian topik merupakan sebuah proses dalam menganalisis data teks untuk menemukan sebuah topik-topik yang ada pada data teks. Pada era digital saat ini, pendeteksian topik sering digunakan untuk menganalisis topik dan mengelompokkan informasi berdasarkan topiknya. Machine learning membantu proses pendeteksian topik menjadi lebih cepat dan efisien, terutama pada data teks dengan ukuran data yang besar. Salah satu metode machine learning yang dapat digunakan untuk pendeteksian topik adalah metode clustering. Namun karena dimensi data yang tinggi membuat beberapa metode clustering kurang efektif menyelesaikan pendeteksian topik. Untuk mengatasi hal tersebut data yang memiliki ukuran dimensi yang cukup tinggi perlu dilakukan proses reduksi dimensi terlebih dahulu. Improved Deep Embedded Clustering (IDEC) merupakan sebuah metode clustering yang secara bersamaan melakukan reduksi dimensi data dan clustering. Oleh karena itu, pada penelitian ini dilakukan pendeteksian topik dengan metode clustering IDEC. Data yang digunakan pada penelitian ini merupakan data berita online AG News, Yahoo! Answer, dan R2. Namun pada metode IDEC, data teks tidak bisa langsung menerima input berupa data teks. Data teks perlu diubah menjadi vektor representasi yang dapat diterima input. Pada penelitian ini digunakan metode representasi teks Bidirectional Encoder Representation from Transformers (BERT). Data teks mula-mula akan diubah oleh BERT menjadi vektor representasi, setelah itu vektor representasi akan diterima dan dilakukan pendeteksian topik oleh metode IDEC. Kemudian pada proses simulasi dilakukan perbandingan kinerja model IDEC dengan representasi teks BERT dan model IDEC dengan representasi teks TF-IDF. Didapatkan hasil simulasi dari kinerja model IDEC dengan representasi teks BERT memiliki kinerja yang lebih unggul dibandingkan dengan model IDEC dengan representasi teks TF-IDF

.....Topic detection is a process in analyzing text data to find topics that exist in text data. In today's digital era, topic detection is often used to analyze topics and grouping the information by topic. Machine learning helps the topic detection process to be faster and more efficient, especially in text data with large data sizes. One of the machine learning methods that can be used for topic detection is the clustering method. However, because the high data dimensions make some clustering methods less effective in completing topic detection. To overcome this, data that has a sufficiently high dimension size needs to be carried out in a dimension reduction process first. Improved Deep Embedded Clustering (IDEC) is a clustering method that simultaneously performs data dimension reduction and clustering. Therefore, in this study, topic detection was carried out using the IDEC clustering method. The data used in this study is the online news data of AG News, Yahoo! Answer, and R2. However, in the IDEC method, text data cannot directly receive input in the form of text data. Text data needs to be converted into a vector representation that can accept input. In this study, the Bidirectional Encoder Representation from Transformers (BERT) text representation method was used. The text data will first be converted by BERT into a vector representation, after that the vector representation will be accepted and topic detection will be carried out by the IDEC method. Then the simulation process compares the performance of the IDEC model with the BERT text representation and the

IDEC model with the TF-IDF text representation. The simulation results obtained from the performance of the IDEC model with the text representation of BERT which has superior performance compared to the IDEC model with the text representation of TF-IDF.