Analisis Sensitivitas Parameter Model EFCM Berbasis BERT untuk Pendeteksian Topik = Parameter Sensitivity Analysis of BERT-based EFCM Model for Topic Detection

Yudhistira Jinawi Agung, author

Abstrak

Pendeteksian topik adalah suatu proses untuk mendapatkan pokok bahasan atau topik pada suatu dokumen teks. Pada data yang besar, pendeteksian topik dapat dilakukan dengan lebih efisien menggunakan metode machine learning. Clustering merupakan salah satu metode machine learning yang bertujuan untuk mengelompokkan data yang memiliki karakteristik serupa ke dalam suatu kelompok/cluster. Beberapa contoh metode clustering adalah K-Means, Fuzzy C-Means (FCM), dan Eigenspace-Based Fuzzy C-Means (EFCM). Metode clustering hanya memproses data numerik, oleh sebab itu diperlukan metode representasi teks. Metode representasi teks yang umum digunakan sebelumnya adalah Bag of Words (BoW) dan Term-Frequency Inversed Document Frequency (TFIDF). Namun, metode BoW dan TFIDF kurang baik dalam merepresentasikan teks secara kontekstual. Pada tahun 2018 metode representasi teks yang baru ditemukan yaitu metode Bidirectional Encoder Representation from Transformers (BERT). Model BERT dapat merepresentasikan teks secara kontekstual dan menghasilkan representasi teks berdimensi tinggi. EFCM merupakan teknik clustering yang menggunakan kombinasi teknik reduksi dimensi Truncated Singular Value Decomposition (TSVD) dengan teknik clustering FCM. Pada tahun 2022 terdapat penelitian yang mengombinasikan BERT dan EFCM untuk pendeteksian topik. Pada model kombinasi BERT dan EFCM terdapat beberapa nilai parameter yang dapat diatur, antara lain adalah pemilihan lapisan encoder BERT, dimensi EFCM, dan derajat fuzziness. Penelitian ini berfokus pada analisis sensitivitas parameter untuk melihat pengaruh dari nilai parameter terhadap kinerja model EFCM berbasis BERT untuk pendeteksian topik. Analisis sensitivitas parameter menggunakan metode Sobol untuk menentukan parameter yang tidak sensitif dan yang paling sensitif. Kinerja model dievaluasi menggunakan metrik evaluasi topic coherence, topic diversity, dan topic quality. Hasil penelitian menunjukkan bahwa parameter lapisan encoder, dimensi EFCM, dan derajat fuzziness sensitif terhadap kinerja model. Selain itu, diperoleh model optimal pada tiga dataset menggunakan parameter tuning metode grid search. Penerapan parameter tuning dapat meningkatkan performa model pada ketiga dataset berdasarkan nilai topic quality.

.....

Topic detection is a process to get the subject matter or topic in a text document. In large data, topic detection can be done more efficiently using machine learning methods. Clustering is a machine learning method aiming to group data with similar characteristics into a group/cluster. Some examples of clustering methods are K-Means, Fuzzy C-Means (FCM), and Eigenspace-Based Fuzzy C-Means (EFCM). The clustering method only processes numeric data; therefore, a text representation method is needed. Previously used text representation methods were Bag of Words (BoW) and Term-Frequency Inverse Document Frequency (TFIDF). However, the BoW and TFIDF methods are not good at representing text contextually.

In 2018 a new text representation method was discovered, namely the Bidirectional Encoder Representation from Transformers (BERT) method. The BERT model can contextually represent text and produce high-dimensional text representations. EFCM is a clustering technique that combines the Truncated Singular Value Decomposition (TSVD) dimension reduction technique with the FCM clustering technique. In 2022 there will be research that combines BERT and EFCM for topic detection. In the BERT and EFCM combination model, there are several parameter values that can be set, including the selection of the BERT encoder layer, EFCM dimensions, and the degree of fuzziness. This study focuses on parameter sensitivity analysis to see the effect of parameter values on the performance of the BERT-based EFCM model for topic detection. Parameter sensitivity analysis uses the Sobol method to determine which parameters are insensitive and the most sensitive. Model performance was evaluated using evaluation metrics of topic coherence, topic diversity, and topic quality. The results showed that the parameters of the encoder layer, EFCM dimensions, and degree of fuzziness were sensitive to model performance. In addition, the optimal model was obtained for three datasets using the grid search method parameter tuning. Parameter tuning can improve the model performance on the three datasets based on topic quality values.