INTISARI

PENGELOMPOKAN TEKS MENGGUNAKAN COSINE SIMILARITY DAN ALGORITMA K-MEDOIDS CLUSTERING

Andika Rahim Darusalam
NIM. 21/475605/PA/06133

Pengelompokan manual jawaban terbuka dari responden survei merupakan proses yang lambat dan tidak efisien, sehingga menghambat analisis kualitatif. Pendekatan yang bisa dilakukan adalah dengan metode *text similarity* tetapi metode ini tidak dapat menangani jawaban di luar kategori yang telah ditentukan. Sementara metode *text clustering* sering kali dipengaruhi oleh data *noise* dan *outlier* sehingga cluster yang terbentuk tidak akurat dan perlu dilakukan analisis manual untuk memisahkan antara *cluster* yang benar dengan *noise* dan *outlier*-nya.

Untuk mengatasi masalah tersebut, penelitian ini mengembangkan sistem pengelompokan otomatis. Solusi ini mengintegrasikan *text similarity* dan *K-Medoids Clustering* yang telah dimodifikasi untuk memfilter *noise* dan *outlier* dengan menetapkan sebuah *threshold*. Proses ini akan membentuk *cluster-cluster* yang lebih akurat dan saling relevan antar anggota *cluster*.

Hasil evaluasi secara konsisten menunjukan keunggulan metode yang diusulkan dibandingkan metode pembandingnya seperti *K-Means, K-Medoids* dan DBSCAN setelah data *noise* dan *outlier* berhasil di-*filter*. Tetapi penggunaan *text similarity* diawal ternyata tidak memberikan dampak besar terlihat dari nilai metrik ARI, AMI, *Purity,* V-Measure dan *Silhouette Score* yang tidak jauh berbeda dengan tanpa menggunakan *text similarity*. Tetapi dengan *text similarity* lebih banyak data yang ter-*cluster*. Dibandingkan DBSCAN, hasilnya lebih unggul metode yang diusulkan dilihat dari metrik evaluasi yang digunakan tetapi DBSCAN lebih banyak berhasil meng-*cluster* data dan lebih sedikit data terdeteksi sebagai *noise* dan *outlier.*

Kata Kunci: ***Text clustering*****,** ***K-Medoids*****, Jawaban Responden,** ***Text similarity*****,** ***General Text Embedding (GTE), Cosine Similarity***

ABSTRACT

TEXT CLUSTERING USING COSINE SIMILARITY AND K-MEDOIDS CLUSTERING ALGORITHM

Andika Rahim Darusalam
NIM. 21/475605/PA/06133

It takes a long time and isn't very effective to manually group open-ended survey responses, which can slow down qualitative analysis. Text similarity can be helpful, but it only works with pre-defined categories and can't handle answers that aren't what you expected. Text clustering methods, on the other hand, often have trouble with noise and outliers, which can cause groups to be wrong and still require manual work to separate the valid clusters from the noise.

To address these challenges, this research introduces an automatic grouping system. The solution combines text similarity with a modified K-Medoids clustering algorithm, which is specifically designed to filter out noise and outliers by using a set threshold. This process aims to create clusters where the members are more accurate and highly relevant to one another.

The evaluation consistently showed that the proposed method outperforms other techniques like K-Means, K-Medoids, and DBSCAN, especially after noise and outliers were filtered out. Interestingly, including text similarity at the beginning didn't significantly boost the quality, as seen in the Silhouette, ARI, AMI, V-Measure and Purity scores, though it did manage to group a larger portion of the data. When compared to DBSCAN, the proposed method also achieved superior cluster quality as seen in the metrics evaluation, but it's worth noting that DBSCAN was successful in clustering a greater number of data points overall.

Keywords: ***Text clustering*****,** ***K-Medoids*****,** ***Respondent Answer*****,** ***Text similarity*****,** ***General Text Embedding (GTE),*** **Cosine Similarity,** ***Silhouette Score***