ABSTRAK

IMPLEMENTASI *CLUSTERING* UNTUK PENGELOMPOKKAN KARYA ILMIAH BERTEMA COVID-19 BERDASARKAN KEMIRIPAN ABSTRAK MENGGUNAKAN *K-MEANS CLUSTERING*

Muhammad Akbar Iskandar

Clustering adalah salah satu proses dalam *data mining* yang digunakan untuk mengelompokkan data dengan data yang lain yang memiliki kemiripan berdasarkan fitur yang dimiliki setiap data untuk dapat memunculkan pola-pola yang ada dalam data yang tidak dapat ditafsirkan secara langsung. Penelitian ini bertujuan untuk melakukan proses *clustering* terhadap abstrak dari 1000 karya ilmiah bertema COVID-19 yang ada pada situs Garuda menggunakan algoritma *K-Means Clustering* dan mencari jumlah *cluster* yang optimal. Transformasi data terhadap abstrak dari karya ilmiah dilakukan menggunakan TF-IDF dan *Bag-of-words*, lalu dilanjutkan dengan metode pengurangan dimensi PCA dan LDA untuk kemudian membandingkan hasil proses *clustering* menggunakan kedua metode pengurangan dimensi tersebut beserta variasi metode pengukuran jarak *euclidean distance* dan *cosine distance*. Hasil evaluasi menggunakan *silhouette coefficient* menunjukkan bahwa hasil *clustering* terbaik dihasilkan dari penggunaan metode pengurangan dimensi LDA dengan penggunaan metode pengukuran jarak *cosine distance* dengan nilai *silhouette* sejumlah 0.5336 untuk 15 *cluster*.

Kata Kunci: *Clustering, Karya Ilmiah, COVID-19, Data Teks, K-Means Clustering, Principle Component Analysis, Latent Dirichlet Allocation*

ABSTRACT

CLUSTERING IMPLEMENTATION FOR GROUPING COVID-19 THEMED SCIENTIFIC PAPERS BASED ON ABSTRACT SIMILARITIES USING K-MEANS CLUSTERING

Muhammad Akbar Iskandar

Clustering is one of the process in data mining that is used to group data with other data that has similarities based on the features of each data to reveal patterns in data that cannot be interpreted directly. This study aims to carry out a clustering process of abstracts from 1000 scientific papers with the theme of COVID-19 on the Garuda website using the K-Means Clustering algorithm and find the optimal number of clusters. Data transformation applied on abstracts from scientific papers is carried out using TF-IDF and Bag-of-words, then proceeded with the PCA and LDA dimension reduction methods, and then compared the results of the clustering process using the two dimension reduction methods along with variations in the Euclidean and Cosine distance measurement methods. The evaluation results using the silhouette coefficient show that the best clustering result are obtained from the LDA dimension reduction method using the cosine distance measurement method with a silhouette value of 0.5336 for 15 clusters.

Keywords**:** *Clustering, Scientific Paper, COVID-19, Textual Data, K-Means Clustering, Principle Component Analysis, Latent Dirichlet Allocation*