ABSTRAK

Nama : Rizal Taufik Rifaldi

NIM : 1157010057

Judul : Perbandingan Performa Metode Deteksi *Outlier Distance-Based* dan *Cluster-Based* pada Algoritma *K-Means* dengan Data Teks Hadis

Outlier adalah objek data yang menyimpang secara signifikan dari objek lainnya. Penelitian ini bertujuan untuk mengidentifikasi dan menghapus *outlier* agar hasil metode *k-means* menjadi lebih baik dari sebelum *outlier* dihapus. *K-means* merupakan algoritma pengelompokkan yang sangat sensitif terhadap *outlier*, sehingga menghapus *outlier* akan mempengaruhi hasil pengelompokkan menjadi lebih baik. *K-means* bekerja dengan cara mengelompokkan data yang memiliki kesamaan berdasarkan jarak antara titik-titik data. Intinya algoritma ini mencoba menemukan pusat dari setiap kelompok dengan cara menghitung jarak antara setiap titik data yang dalam kasus ini menggunakan pengukuran jarak *cosine similarity* dan *euclidean distance*. Selain itu pusat kelompok tersebut dihitung sebagai rata-rata dari semua titik data di dalam kelompok. Kemudian hasil dari pengelompokkan tersebut dievaluasi menggunakan *silhouette coefficient.* Beberapa teknik deteksi *outlier* yang digunakan untuk algoritma *k-means* adalah *distance based* dan *cluster based*. Di mana *distance based* menentukan *outlier* berdasarkan nilai *threshold*, jika suatu titik data jauh dari pusat *cluster* maka dianggap sebagai *outlier*, sementara *cluster based* menentukan *outlier* berdasarkan *cluster* yang memiliki anggota yang paling sedikit. Data yang dijadikan sebagai objek penelitian adalah data teks hadis dan memiliki 5 topik pembahasan yaitu, tafsir Al-Qur'an, peperangan, perilaku yang terpuji, haji dan jum'at. Di mana yang menjadi *input* adalah data hadis yang telah direduksi fitur menggunakan metode *principal component analysis* (PCA) dan data hasil pembobotan kata menggunakan metode *Term frequency – Inverse document frequency* (TF-IDF). Hasil dari penelitian menyimpulkan bahwa menggunakan data hasil reduksi fitur *principal component analysis* (PCA) lebih baik dari segi kualitas cluster maupun *runtime*-nya dari menggunakan data hasil pembobotan kata *Term frequency – Inverse document frequency* (TF-IDF). Lalu pengukuran jarak pada algoritma *k-means* menggunakan *cosine similarity* lebih baik dari pada *euclidean distance*. Kemudian metode deteksi *outlier* menggunakan *cluster based* lebih baik dari *distance based* dalam segi performa pengelompokkan yang mengacu pada evaluasi *cluster*.

Kata Kunci: *Outlier*, Deteksi *Outlier, K-Means, Distance Based, Cluster Based, Silhouette Coefficient, Principal Component Analysis*

ABSTRACT

Name **: Rizal Taufik Rifaldi**

NIM **: 1157010057**

Title **: Perbandingan Performa Metode Deteksi *Outlier Distance-Based* dan *Cluster-Based* pada Algoritma *K-Means* dengan Data Teks Hadis**

Outliers are data objects that deviate significantly from other objects. This study aims to identify and remove outliers in order to improve the results of the k-means method compared to when the outliers were not removed. K-means is a clustering algorithm that is highly sensitive to outliers, so removing outliers can have a positive impact on the clustering results. K-means works by grouping data points that are similar based on the distance between them. Essentially, this algorithm attempts to find the center of each group by calculating the distance between each data point, using similarity measures such as cosine distance and euclidean distance. The center of each group is then calculated as the average of all data points within that group. The results of the clustering are evaluated using silhouette coefficients. Several outliers detection techniques can be used in conjunction with the k-means algorithm, including distance-based and cluster-based methods. Distance-based techniques determine outliers based on a threshold value, considering a data point as an outlier if it is far from the center of its cluster. On the other hand, cluster-based techniques determine outliers based on the cluster with fewest members. For this study, hadith text data is used as the research object, consisting of five discussion topics: interpretation of the Qur'an, war, commendable behavior, pilgrimage and Friday. The input data has been preprocessed by applying principal component analysis (PCA) for feature reduction and term frequency-inverse document frequency (TF-IDF) weighting for word representation. The study results concluded that using PCA for feature reduction yielded better cluster quality and runtime compared to using TF-IDF weighted data. additionally, distance measurement using Euclidean distance. Lastly, the cluster-based outlier detection method outperformed the distance-based method in terms of grouping performance, as indicated by cluster evaluation metrics.

Keywords: *Outlier*, Deteksi *Outlier*, *K-Means*, *Distance Based*, *Cluster Based*, *Silhouette Coefficient*