

ABSTRAK

Nama : Deni Kurniawan

NIM : 1207010014

Judul : Relasi Semantik Kata pada Data Terjemahan Al-Qur'an Bahasa Indonesia Menggunakan Algoritma *Clustering K-means* Bertingkat

Kemajuan teknologi informasi telah menghasilkan volume data yang sangat besar dan beragam. Untuk mengekstrak informasi dari data ini, dilakukan penggalian data (*data mining*). Dalam konteks data teks, objek yang diekstraksi bisa berupa kata, kalimat, paragraf, atau dokumen. Metode-metode data mining berbasis *embedding* prediktif masih jarang diterapkan pada dokumen-dokumen Islam, yang menjadi motivasi utama penelitian ini. Penelitian ini menggunakan *word2vec* untuk menghasilkan representasi numerik dari kata-kata dalam teks terjemahan Al-Qur'an bahasa Indonesia. Metode *clustering* digunakan untuk mengelompokkan representasi numerik dari kata-kata ini. Dengan menerapkan pendekatan hierarki pada algoritma *clustering k-means* bertingkat, data dikelompokkan secara bertahap melalui beberapa iterasi *clustering*, yang memungkinkan pembentukan *subcluster* di dalam *cluster* utama. Proses ini tidak hanya mempermudah penentuan nilai *k* secara bertahap, tetapi juga membantu dalam menyeimbangkan ukuran *cluster* dengan lebih baik. Analisis hasil *clustering* menunjukkan bahwa kelompok-kelompok kata berdasarkan logika linguistik yang diperoleh dari algoritma *clustering k-means* bertingkat memiliki nilai *cosine similarity* yang tinggi yang mendekati nilai 1, hal ini menunjukkan keterikatan semantik yang kuat antar kata dalam *cluster*. Pola karakteristik relasi semantik kata dari *cluster* yang diperoleh mencerminkan frekuensi dan kemunculan kata-kata dalam ayat-ayat Al-Qur'an pada konteks topik yang sama, di mana kata-kata yang sering muncul bersama dalam konteks yang sama akan memiliki vektor yang lebih dekat satu sama lain.

Kata Kunci : *Data Mining, Word2vec, K-means Bertingkat, Silhouette Coefficient, Cosine Similarity*

ABSTRACT

Name : Deni Kurniawan

NIM : 1207010014

Title : *Semantic Relationships of Words in Indonesian Al-Qur'an Text Data Using the Leveling K-means Clustering Algorithm*

Advances in information technology have produced very large and varied data volumes. To extract information from this data, data mining is carried out. In the context of text data, the objects extracted can be words, sentences, paragraphs, or documents. Data mining methods based on predictive embedding are still rarely applied to Islamic documents, which is the main motivation for this research. This research uses word2vec to produce numerical representations of words in the Indonesian translation of the Al-Qur'an. Clustering methods are used to group the numerical representations of these words. By applying a hierarchical approach to the multilevel k-means clustering algorithm, data is grouped gradually through several clustering iterations, which allows the formation of subclusters within the main cluster. This process not only makes it easier to determine the value of k step by step, but also helps in better balancing the cluster sizes. Analysis of the clustering results shows that groups of words based on linguistic logic obtained from the multilevel k-means clustering algorithm have high cosine similarity values which are close to 1, this indicates a strong semantic attachment between words in the cluster. The characteristic pattern of word semantic relations from the clusters obtained reflects the frequency and occurrence of words in the verses of the Qur'an in the context of the same topic, where words that often appear together in the same context will have closer vectors one another.

Keyword : Data Mining, Word2vec, Leveling K-means, Silhouette Coefficient, Cosine Similarity