ABSTRAK

Nama : Mochamad Rajib Deyana

NIM : 1137010037

Judul Skripsi : *Document Embedding* Menggunakan *Paragrah Vector* untuk *Clustering* Terjemahan Ayat-ayat Al-Qur'an

Perkembangan pesat dalam bidang teknologi informasi menghasilkan data yang sangat besar dan beragam. Penggalian data (*data mining*) kemudian dilakukan untuk mengekstrak informasi yang berguna. Ketika dihadapkan pada data teks, objek dapat berupa kata, kalimat, paragraf, maupun dokumen, di mana *text embedding* digunakan untuk mengonversi objek tersebut ke dalam bentuk numerik. Masih jarangnya penerapan metode-metode *data mining* menggunakan pendekatan *embedding* berbasis prediksi pada naskah-naskah keagamaan menjadi motivasi utama dari penelitian ini. *Paragraph vector* kemudian digunakan untuk menghasilkan sebuah representasi numerik dari teks berupa terjemahan dan tafsir-tafsir Al-Qur'an dalam Bahasa Indonesia. Dari tes analogi berupa 173.002 pertanyaan semantik dan 208.920 pertanyaan sintaksis yang diberikan, vektor kata yang dihasilkan *paragraph vector* mampu menjawab masing-masing sebesar 47,04% dan 56,75% pertanyaan dengan benar. Metode *data mining* seperti *clustering* selanjutnya digunakan untuk mengelompokkan vektor dokumen dari terjemahan pokok-pokok bahasan Al-Qur'an. Menggunakan CLARANS, diperoleh 8 kelompok pokok bahasan Al-Qur'an yang berkorelasi dengan nilai terbaik pada pengukuran internal *cluster*: *Silhouette Coefficient* dan *Davies-Bouldin Index* masing-masing sebesar 0,0965 dan 1,8038.

Kata Kunci: *data mining*, *text embedding*, tafsir Al-Qur'an, *word2vec*, *paragraph vector*, tes analogi, *clustering*, CLARANS, pengukuran internal *cluster*

ABSTRACT

Name : Mochamad Rajib Deyana

NIM : 1137010037

Title : *Document Embedding* using *Paragrah Vector* for *Clustering* Translations of Verses of the Qur'an

The rapid development of information technology produces very large and diverse data. *Data mining* is then carried out to extract useful information. When dealing with text data, objects can be words, sentences, paragraphs, or documents, where *text embedding* is used to convert these objects into numeric form. The lack of application of *data mining* methods in *predictive-based text embedding* for religious texts is the motivation of this study. *Word2vec* and *paragraph vector* are then used to produce a numerical representation of the translations and interpretations (*Tafseer*) of the Qur'an in Indonesian. From the analogy test in the form of 173,002 semantic questions and 208,920 syntactic questions, word vectors obtained by *paragraph vector* is able to answer 47.04% and 56.75% of each question type correctly. *Data mining* methods such as *clustering* are then used to classify document vectors from the translation of the subjects of the Qur'an. Using CLARANS, 8 groups of Al-Qur'an subjects were obtained with the best value in the internal cluster measurements: *Silhouette Coefficient* and *Davies-Bouldin Index* of 0.0965 and 1.8038, respectively.

Keyword: *data mining, text embedding, Tafseer, word2vec, paragraph vector*, analogy test, *clustering*, CLARANS, internal cluster measurement