ABSTRAK

Nama : **Maulidah Tsaniatuluzzma**

NIM : **1217010042**

Judul Skripsi : **Pengaruh Hyperarameter Dalam FastText Terhadap *Semantic Similarity* Kata Menggunakan Dataset Al-Qur'an Bahasa Arab**

Penelitian ini menganalisis pengaruh hyperparameter FastText terhadap kemiripan semantik kata pada dataset Al-Qur'an berbahasa Arab. Bahasa Arab Al-Qur'an yang kompleks memerlukan metode representasi kata yang akurat. Metodologi mencakup pengumpulan dataset 6.236 ayat dari Tanzil.net, diikuti pra-pemrosesan seperti tokenisasi, pembersihan teks, penghapusan *stopwords*, normalisasi, dan *lemmatisasi*. Model FastText dilatih menggunakan arsitektur Skip-gram, yang efektif untuk kata jarang muncul. Hyperparameter yang diuji meliputi dimensi vektor, *window size*, *learning rate*, *minimum count*, dan epoch. Evaluasi *semantic Similarity* menggunakan *cosine Similarity* menunjukkan konfigurasi optimal adalah dimensi *embedding* 300, 10 epoch, dan *window size* 5, dengan nilai similaritas mencapai 0.999. Visualisasi PCA memvalidasi kemampuan model FastText mengelompokkan kata berdasarkan makna dan konteks dalam Al-Qur'an, termasuk hubungan sinonim.

Kata Kunci: FastText, Hyperparameter, *Semantic Similarity*, Al-Qur'an Bahasa Arab, *Word Embedding*.

ABSTRACT

Nama : **Maulidah Tsaniatuluzzma**

NIM : **1217010042**

Judul Skripsi : **The Effect of Hyperparameters in FastText on Semantic Similarity of Words Using the Arabic Qur'an Dataset.**

This study analyzes the impact of FastText hyperparameters on word semantic Similarity using an Arabic Quranic dataset. The complex nature of Quranic Arabic necessitates accurate word representation methods. The methodology involved collecting a 6,236-verse dataset from Tanzil.net, followed by pre-processing steps such as tokenization, text cleaning, stopword removal, normalization, and lemmatization. The FastText model was trained using the Skip-gram architecture, which is effective for rare words. Tested hyperparameters included vector dimension, window size, learning rate, minimum count, and epochs. Semantic Similarity evaluation using cosine Similarity showed optimal configuration with an embedding dimension of 300, 10 epochs, and a window size of 5, achieving a Similarity value of 0.999. PCA visualization validated FastText's ability to cluster words based on meaning and context within the Quran, including synonym relationships.

Keywords: *FastText, Hyperparameters, Semantic Similarity, Arabic Quran, Word Embedding.*