

ABSTRAK

Nama : Ita Pebriyanti

NIM : 1227010029

Judul : **OPTIMISASI DIMENSI *WORD EMBEDDINGS* BAHASA ARAB
MENGUNAKAN METODE *PAIRWISE INNER PRODUCT* (PIP)
LOSS BERBASIS TEORI PERTURBASI MATRIKS**

Dimensi *word embedding* merupakan salah satu parameter yang sangat memengaruhi kualitas representasi kata, namun penentuannya hingga saat ini masih banyak dilakukan secara heuristik. Penelitian ini bertujuan untuk menentukan dimensi optimal *word embedding* bahasa Arab secara analitis menggunakan pendekatan *Pairwise Inner Product* (PIP) *Loss* yang berlandaskan Teori Perturbasi Matriks. Tiga korpus Wikipedia bahasa Arab yang berbeda ukuran dan topik domainnya digunakan sebagai data penelitian. Matriks *Pointwise Mutual Information* (PMI) didekomposisi melalui *Singular Value Decomposition* (SVD), kemudian tingkat *noise* diestimasi menggunakan *Count-Twice Trick* dan disaring melalui *Universal Singular Value Thresholding* (USVT). Dimensi optimal diperoleh berdasarkan Teorema 3 dan divalidasi secara empiris melalui Simulasi Monte Carlo. Evaluasi kualitas representasi kata dilakukan menggunakan dataset *WordSim-353* dengan korelasi peringkat Spearman. Hasil penelitian memperlihatkan bahwa karakteristik korpus berpengaruh signifikan terhadap tingkat *noise* dan jumlah dimensi efektif yang dihasilkan. Pada korpus berskala besar, Teorema 3 dan Monte Carlo menghasilkan dimensi optimal yang hampir identik, sedangkan pada korpus berukuran kecil selisihnya melebar. Dalam evaluasi intrinsik, Teorema 3 menunjukkan keunggulan atau kinerja setara dengan Monte Carlo pada dua dari tiga korpus yang diuji. Berdasarkan temuan tersebut, dapat disimpulkan bahwa Teorema 3 merupakan metode penentuan dimensi optimal yang efisien dan memiliki landasan teoritis yang kuat.

Kata Kunci *word embedding*, dimensi optimal, PIP *Loss*, teori perturbasi matriks, USVT, bahasa Arab

ABSTRACT

Name : Ita Pebriyanti

NIM : 1227010029

Title : *OPTIMIZATION OF ARABIC WORD EMBEDDINGS DIMENSIONS USING PAIRWISE INNER PRODUCT (PIP) LOSS METHOD BASED ON MATRIX PERTURBATION THEORY*

Word embedding dimension is a critical parameter that significantly affects the quality of word representations, yet its selection remains largely heuristic. This study aims to analytically determine the optimal dimensions of Arabic word embeddings using the Pairwise Inner Product (PIP) Loss approach grounded in Matrix Perturbation Theory. Three Arabic Wikipedia corpora differing in size and domain topic were employed as research data. The Pointwise Mutual Information (PMI) matrix was decomposed via Singular Value Decomposition (SVD), after which the noise level was estimated using the Count-Twice Trick and filtered through Universal Singular Value Thresholding (USVT). The optimal dimension was obtained based on Theorem 3 and empirically validated through Monte Carlo Simulation. Word representation quality was evaluated using the WordSim-353 dataset with Spearman's rank correlation. The results reveal that corpus characteristics significantly influence the noise level and the number of effective dimensions. For the large-scale corpus, Theorem 3 and Monte Carlo produced nearly identical optimal dimensions, whereas the gap widened for smaller corpora. In intrinsic evaluation, Theorem 3 demonstrated superiority or equivalent performance to Monte Carlo in two out of three corpora tested. Based on these findings, it can be concluded that Theorem 3 is an efficient method for determining optimal dimensions with a strong theoretical foundation.

Keywords word embedding, optimal dimension, (PIP) Loss, matrix perturbation theory, USVT, Arabic language