

ABSTRAK

Nama : Isma Kusmiati

NIM : 1227010028

Judul : *Neural Probabilistic Language Model (NPLM) Modern* pada Korpus Terjemahan dan Tafsir Al-Qur'an Bahasa Indonesia

Model bahasa statistik berbasis *n-gram* memiliki keterbatasan berupa *curse of dimensionality* dan *data sparsity*, sedangkan *Neural Probabilistic Language Model (NPLM)* awal masih menggunakan jaringan dangkal dan konteks lokal. Penelitian ini menerapkan dan mengevaluasi NPLM modern berdasarkan gagasan *revisiting NPLM* dari Sun dan Iyyer (2021). Model yang digunakan mencakup jaringan *feed-forward residual*, *layer normalization*, *adaptive softmax*, dan agregasi konteks global. Eksperimen dilakukan pada korpus terjemahan dan tafsir Al-Qur'an berbahasa Indonesia yang terdiri atas 6.236 data dan kosakata sebanyak 12.184 kata. Konfigurasi pelatihan ditetapkan melalui analisis sensitivitas *One-Factor-at-a-Time (OFAT)* dan eksperimen kombinasi naif. Studi ablasi membandingkan tiga arsitektur, yaitu NPLM tanpa konteks global, NPLM dengan *average pooling*, dan NPLM dengan *kernel aggregator*. Konfigurasi *Global Context Kernel* menghasilkan *validation perplexity* terendah, yaitu 81,81, dibandingkan dengan *Average Pooling* sebesar 83,39 dan model dasar sebesar 91,10. Hasil tersebut menunjukkan penurunan sekitar 9,29. Analisis probabilitas menunjukkan entropi lebih rendah pada konteks dengan frasa yang lebih tetap dan lebih tinggi pada konteks naratif terbuka. Representasi dengan *cosine similarity* dan *nearest neighbor* menunjukkan pola kedekatan distribusional pada sejumlah kata yang saling berkaitan. Hasil penelitian menunjukkan bahwa NPLM modern, khususnya dengan *kernel aggregator*, meningkatkan kinerja pemodelan teks keagamaan berbahasa Indonesia.

Kata Kunci: *Neural Probabilistic Language Model, revisiting NPLM, konteks global, kernel aggregator, tafsir Al-Qur'an, validation perplexity, word embedding*

ABSTRACT

Name : Isma Kusmiati
Student ID : 1227010028
Title : *Neural Probabilistic Language Model (NPLM) modern on an Indonesian Qur'an Translation and Tafsir Corpus*

Statistical n-gram language models are limited by the curse of dimensionality and data sparsity, while early Neural Probabilistic Language Models (NPLMs) rely on shallow networks and local contexts. This study implements and evaluates a modernized NPLM based on the revisiting NPLM framework proposed by Sun and Iyyer (2021). The model incorporates residual feed-forward networks, layer normalization, adaptive softmax, and global context aggregation. Experiments were conducted on an Indonesian corpus of Qur'anic translations and interpretations containing 6,236 data records and a vocabulary of 12,184 words. Training configurations were determined through One-Factor-at-a-Time (OFAT) sensitivity analysis and naive combination experiments. An ablation study compared three architectures: NPLM without global context, NPLM with average pooling, and NPLM with a kernel aggregator. The Global Context Kernel configuration achieved the lowest validation perplexity at 81.81, compared with 83.39 for Average Pooling and 91.10 for the baseline, representing an approximately 9.29 reduction. Probability analysis showed lower entropy in more fixed phrase contexts and higher entropy in open-ended contexts. Cosine similarity and nearest-neighbor analyses indicated distributional proximity among selected related words. These findings suggest that the modernized NPLM, particularly with a kernel aggregator, improves language modeling performance on Indonesian religious texts.

Keywords: *Neural Probabilistic Language Model, revisiting NPLM, global context, kernel aggregator, Qur'anic interpretations, validation perplexity, word embedding*