

ABSTRAK

Kemampuan membaca merupakan keterampilan literasi dasar yang sangat penting bagi perkembangan anak, namun anak dengan disleksia memiliki kesulitan dalam membaca dan memahami teks bacaan karena penggunaan kosakata yang sulit, abstrak, dan panjang. Penelitian ini bertujuan untuk mengembangkan sebuah metode simplifikasi kosakata Bahasa Indonesia guna mendukung literasi baca anak disleksia usia 7-12 tahun. Pendekatan yang diusulkan yaitu kombinasi antara analisis statistik berbasis frekuensi kata dengan algoritma *clustering* K-Means untuk mengidentifikasi tingkat kesulitan kata. Frekuensi kata digunakan sebagai indikator familiaritas kosakata, sementara algoritma K-Means mengelompokkan kata ke dalam kluster “mudah” dan “sulit” secara *unsupervised*. Terhadap kosakata yang teridentifikasi sulit, sistem menggunakan GPT-3.5 Turbo untuk memberikan rekomendasi kata pengganti yang lebih sederhana tanpa mengubah makna utama teks. Metodologi penelitian yang digunakan yaitu *Knowledge Discovery in Database* (KDD) yang meliputi tahap *selection*, *preprocessing*, *transformation*, *data mining*, dan *evaluation*. Hasil penelitian menunjukkan bahwa K-Means mampu mengelompokkan 3.695 kosakata ke dalam dua kluster dengan nilai *Silhouette Score* sebesar 0,7594 dan Davies–Bouldin Index sebesar 0,3626. Pada proses rekomendasi, sebanyak 1.510 kata (63,1%) menghasilkan kata pengganti yang lebih pendek dan lebih mudah dipahami, dengan sebagian besar rekomendasi memiliki nilai *semantic similarity* di atas 0,70. Hasil tersebut menunjukkan bahwa metode yang dikembangkan mampu mengidentifikasi kosakata sulit dan menghasilkan rekomendasi kata yang lebih sederhana dengan tetap mempertahankan keterkaitan makna.

Kata Kunci: Disleksia, Frekuensi Kata, K-Means, *Natural Language Processing*, OpenAI, Simplifikasi Kosakata.

ABSTRACT

Reading ability is a fundamental literacy skill that is very important for a child's development, but children with dyslexia have difficulty reading and understanding texts due to the use of difficult, abstract, and lengthy vocabulary. This research aims to develop a method for simplifying Indonesian vocabulary to support reading literacy in children with dyslexia aged 7-12 years. The proposed approach is a combination of frequency-based statistical analysis of words with the K-Means clustering algorithm to identify the difficulty level of words. Word frequency is used as an indicator of vocabulary familiarity, while the K-Means algorithm clusters words into "easy" and "difficult" categories in an unsupervised manner. For the vocabulary identified as difficult, the system uses GPT-3.5 Turbo to provide recommendations for simpler substitute words without changing the main meaning of the text. The research methodology used is Knowledge Discovery in Database (KDD), which includes the stages of selection, preprocessing, transformation, data mining, and evaluation. The research results show that K-Means is capable of clustering 3,695 vocabulary words into two clusters with a Silhouette Score of 0,7594 and a Davies–Bouldin Index of 0,3626. In the recommendation process, 1,510 words (63.1%) resulted in shorter and more easily understood substitutes, with most recommendations having a semantic similarity score above 0,70. These results indicate that the developed method is capable of identifying difficult vocabulary and generating simpler word recommendations while maintaining semantic relevance.

Keywords: *Dyslexia, Word Frequency, K-Means, Natural Language Processing, OpenAI, Vocabulary Simplification.*

