ABSTRACT

Stemming is a processs to find root word from its complex form by removing all affixes are attached on it. Stemming have been applied in text or document clustering, classification, summarization, information retrieval and word-based text compression.

Various language stemmers have been developed, included Indonesian, but Indonesian lamguage is one of the most complicated amongs other languages. Indonesian language has complex affix forms, there are prefixes, infixes, suffixes, confixes, and repeated forms. In Indonesian language, there are morfological change when a root word is attached with affixes particularly prefixes.

The first Indonesian stemmer was developed by Nazief-Adriani then Jelita Asian improved the algorithm called confix stripping (CS) stemmer. There were heaps of improvement was done by CS stemmer so it is highest accuracy stemmer algorithm, but there are still stemming failures.

A new algorithm would be proposed to improve CS stemmer algorithm by modifying algorithm specifically by rearrange stemming process steps sequence. Experiment would be performed to compare the accuracy amongs Nazief – Adriani, CS stemmer, and new algorithm by using all of those algorithm to stemm the words from 3 document sources, those were a novel book, a hadits book, and online news. Stemming processses used a root word dictionary parsed from “Kamus Besar Bahasa Indonesia 2008”. Result of experiment showed that new algorithm have better accuracy than both Nazief-Adriani and CS stemmer.

Keyword : Stemming, Indonesian, Nazief-Adriani, CS stemmer, new algorithm