ABSTRACT

Lemmatization is a technique in natural language processing that is used to return a word to its basic word based on the Indonesian dictionary. Lemmatization is used on related needs with text mining such as information retrieval which performed at the preprocessing stage.

The method of lemmatization in Indonesian is better known as stemming. In 1996, Nazief had built a stemming system, but there were still some mistakes. Later, many researchers have improved Indonesian lemmatization / stemming system.

Sastrawi and Widayanto are researchers who have built a system of lemmatization that improve the previous algorithm. Both have better accuracy than previous algorithm, but their results are unsuccessful due to morphological rules that are not inputted into the algorithm, this makes the accuracy obtained is still not optimal. In addition, the previous algorithm takes a lot of time to run the algorithm that makes performance slow.

This research has improved Widayanto's algorithm by adding several morphology rules and improving incorrect words caused by typo with addition spellchecker algorithm. The proposed algorithm also rearranges the algorithm to speed up the process of lemmatization. The dataset used in this study is literary novels, religious books, and news with a total of 8 datasets. This study improves the accuracy and performance of the Widayanto's and Sastrawi's methods.

Keywords: Lemmatization, Spellchecker, Indonesian Language, Information Retrieval