ABSTRACT

DETECTION OF PLAGIARISM USING LEXICAL SCORE

By

ANDHIKA SATRIA BAGASKORO
15/379616/PA/16674

Plagiarism often occurs in the academic sphere, especially in higher education institutions, because it is easy to access other people's documents. There are many methods to detect plagiarism and are divided into two parts, namely extrinsic plagiarism detection and intrinsic plagiarism detection. Extrinsic plagiarism detection method is a method that uses a document that is considered plagiarism and is compared with other documents to detect plagiarism. This method requires a lot of documents as a comparison for accurate results. Whereas intrinsic plagiarism detection method can detect using only a few documents.

In this study, the intrinsic plagiarism detection method is used by using a lexical score to see its performance and compared with the method by Kuznetsov et al., 2016. The data used comes from PAN-2018. This method uses word frequency in each sentence, paragraph or document, lexical score, use of word classes and use of punctuation. These features will then be trained using the Gradient Boosting Regressor which generates sentence values. By using a certain threshold value, each sentence will be labeled plagiarism or not.

By using a lexical score, the system can detect plagiarism sentences with the highest F1 score of 42.67%. These results have a significant increase, when compared to not using a lexical score.

Keywords: Plagiarism Intrinsic Detection, Plagiarism, lexical score, Gradient Boosting Regressor, PAN-2018