Abstract

The development of digital information is now developing very rapidly, the information can be news text, the song lyric, or article. The high internet users caused many documents which are not known who or what the gender of the author. Many people use social media to write and make text on internet increase. A question we address in this paper that, can we identify if the author is a man or a woman? And are these people did not faked their gender? Text mining analysis is so important for this problem. If we know who is the author of the text we can decrease plagiarism case that can disadvantage the author of the text.

One of text mining that we can do is classification. There are some methods for classification such as Support Vector Machine, Naive Bayes Classsifier, dan Decision Tree. On previous research [1] to identify gender from a text used three different methods such as Support Vector Machine, Bayesian-based logistic regression and AdaBoost decision tree. Features extraction on this study use some features such as character-based, word-based, syntactic, structure-based, and function words.

Based on the result of the research, the combination of Character-Based Features dan Syntactic-Based Features is the lowest one with 42% accuracy while the combination of Syntactic-Based Features, Structurally-Based Features and Function Word-Based Features is the highest one with 76% accuracy.

Keyword *: document classification, author gender classification, support vector machine.*