Abstract

Internet has many positive impacts in human life. But some of the negative effects also appear due to the freedom of information offered through the internet media. One of the negative effects is easy access to porn sites.

A mechanism is needed to filter porn content that can intelligently identify whether a *web page* contains porn content or not. The mechanism of filtering is discussed in this final project using *naive bayes classifier* method. It is expected that by using *naive bayes classifier* can be obtained high accuracy of the classfication results although each *web page* have different languages, Bahasa Indonesia and English.

In the implementation of *naive bayes classifier* , the *web page* must preprocess-first. *Web page* will be extracted to token / term. The problem appears when there are token / term variants that have a zero value that can cause inaccuracies in the calculation of *posterior probability naive bayes classifier* . Therefor, the author offers three alternatives attribute handling the token / term variants that have a zero value, they are the ignoring the token / term variants that have a zero value, the addition of dummy instance, and initiating the minimum standard deviation. It also analyzed the influence of the use of *stopword removal* on the preprocessing in improving the classification accuracy.

Shown in this final project that the best alternative in handling the token / term variants that have a zero value is initialising with the minimum standard deviation and *stopword removal*. The accuracy of this alternative is relatively high with 96%.

Keywords: *naive bayes classifier* , filtering, porn, preprocess, *stopword removal*, zero variants token