Intisari

Penapisan konten pornografi merupakan tugas yang sangat diperlukan. Sistem penapis berbasis pemblokiran URL yang saat ini diterapkan memiliki banyak kelemahan. Alternatif lain yang dapat dipertimbangkan adalah penapisan pornografi berbasis konten, baik konten tekstual ataupun visual. Dalam banyak penelitian, klasifikasi berbasis konten tekstual memegang peranan penting dalam deteksi awal keberadaan konten porno. Akan tetapi, klasifikasi konten berbasis teks bergantung pada bahasa yang digunakan. Penelitian terkini mengenai penapis teks pornografi berbahasa Indonesia dilakukan dengan penggunaan VSM dan TF-IDF. Namun demikian, akurasi yang dihasilkan perlu ditingkatkan

Penelitian ini berupaya membandingkan dan memilih model klasifikasi yang lebih akurat dari pada penelitian sebelumnya. Akurasi klasifikasi teks dipengaruhi banyak faktor. Faktor yang berpengaruh diantaranya adalah koleksi data, banyak fitur yang digunakan, *corpus category*, metode pra-proses, dan pemilihan algoritme klasifikasi. Penelitian ini menggunakan metode *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC) yang pada banyak penelitian klasifikasi teks menunjukkan performa yang baik. Penelitian ini juga melakukan pengujian berbagai metode pra-proses teks. Penelitian ini menggunakan 200 data latih dan 186 data uji.

Hasil penelitian menunjukkan bahwa akurasi klasifikasi dipengaruhi oleh metode pra-proses yang diterapkan. Hal ini berlaku baik pada metode klasifikasi SVM ataupun NBC. Akurasi klasifikasi tertinggi yang dihasilkan adalah sebesar 97.85%.

Kata kunci: *Machine Learning*, Pornografi, Klasifikasi, *Support Vector Machine* (SVM), *Naïve Bayes Classifier* (NBC), Pra-proses, Teks

Abstract

Pornographic content filtering is an indispensable task. URL-based filtering system that is currently applied has many weaknesses. An alternative approach that can be considered is content-based pornographic filtering using classification algorithm, using textual or visual content. The textual content-based classification plays an important role in early detection of pornographic content. However, it depends on the used language. Recent research on text classification of pornographic content in Indonesian language was done by the use of VSM and TF-IDF. However, the resulting accuracy needs to be improved.

This research objectives are to compare and choose classification model that is more accurate than the previous research. Text classification accuracy is influenced by many factors. They are collection of data, number of features, corpus category, preprocessing method, and the classification algorithm. This study uses the Support Vector Machine (SVM) and Naïve Bayes classifier (NBC) which show good performance in many text classification studies. This research also tests various methods of text preprocessing. This research uses 200 training data and 186 testing data.

The results show that the classification accuracy is influenced by the preprocessing methods applied. This applies both to the SVM and NBC classification. The highest classification accuracy in this research is 97.85%.

Keywords: *Machine Learning, Pornography, Classification, Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), Preprocessing, Text*