ABSTRACT

The growth of porn website in Indonesia has increased, more varied and more accessible. This problem may lead to a bad impact towards the behavior of either children or teenagers. In order to monitor usage, people can use the option of blocking the URL (Uniform Resource Locater), which is also known as URL Blocking. If a URL is blocked, then user will not be able to view the URL address or its web content.

An alternative method to solve this problem is to develop a system that is able to detect websites that contain negative contents. For detecting those websites, a negative content can be determined based on the text contained in the website. In this research, the detection process was performed by existing text contained in the negative website. There are four processes for detecting negative contents, 1). Text extraction with DOM Tree method, 2). Language Detection process with Naïve Bayes algorithm, 3). Classification process with TF-IDF method and Vector Space Model and 4). Negative text removal with Regular Expression (Regex).

The system works such as a proxy in a server, when the user accessed the website pages will be analyzed by the system to determine its category. If the pages category is porn, the system is done the blocking. Blocking of pornographic websites only eliminate negative (porn) text without omit information or other text. Blocking has done by searching for negative words (porn) contained in the website porn's category and replace the porn text with asterisk (*). The accuracy of the system's classification is 82.80% for Indonesian websites and 83.87% for English websites. There are three factors that make misclassification in this system, namely misinterpreted, misspelling and less frequent porn's words.

Keywords—pornography; classification; TF-IDF; *Vector Space Model*

INTISARI

Perkembangan website yang menyediakan konten negatif (porno) di Indonesia semakin banyak, beragam dan mudah untuk diakses. Masalah ini menimbulkan dampak buruk terhadap perilaku anak atau remaja. Salah satu cara untuk mengatsi akses pada website negatif (porno) adalah pemblokiran URL (*Uniform Resource Locater*) atau dikenal dengan URL *Blocking*. Jika sebuah URL diblokir, maka pengguna tidak bisa mengakses alamat URL atau konten yang terdapat dalam website tersebut.

Salah satu cara untuk mengatasi permasalahan ini adalah dengan membuat suatu sistem yang bisa melakukan deteksi terhadap website yang mengandung konten negatif. Pendeteksian website negatif (porno) bekerja berdasarkan teks yang terdapat di dalam halaman website. Proses untuk melakukan pendeteksian dan pemblokiran terdiri dari empat proses, yakni 1) Ekstraksi teks dengan metode DOM *Tree*, 2) Proses *Language Detection* untuk memisahkan website berdasarkan bahasa yang digunakan (Bahasa Inggris atau Bahasa Indonesia) menggunakan algoritme Naïve Bayes, 3) Proses klasifikasi dengan menggunakan metode TF-IDF dan *Vector Space Model* (SVM) dan 4) Proses menghilangkan konten teks negatif dengan metode *Regular Expression* (Regex).

Sistem bekerja yang dibangun bekerja seperti *proxy* di dalam sebuah server. Website yang diakses pengguna akan dianalisa oleh sistem untuk mengetahui kategorinya. Jika website masuk dalam kategori porno, maka sistem akan melakukan pemblokiran dengan cara mencari kata-kata negatif (porno) yang terdapat di dalam website kategori porno dan menggantikannya dengan tanda bintang (*). Dari hasil penelitian menunjukan akurasi sistem dalam mendeteksi website negatif Bahasa Indonesia sebesar 82,80% dan akurasi pendeteksian website negatif Bahasa Inggris sebesar 83,87%. Kesalahan klasifikasi disebabkan oleh tiga hal, yakni konteks kalimat, kesalahan dalam mengeja kata dan frekuensi kemunculan kata porno yang sedikit.

Kata Kunci: Pornografi, Klasifikasi, TF-IDF, *Vector Space Model* (VSM), Regex, Naïve Bayes