

## ABSTRACT

Text-based Indonesian sensitive sites detection and filtering systems so far have not produced good performance to meet the expectation. The system performance—especially in term of accuracy, precision, and sensitivity—can be enhanced substantially because it has not reached the ideal performance of 100%.

This research aims to build a sensitive sites detection and filtering system with higher accuracy, precision, and recall on text classification using a proper classification technique i.e. Support Vector Machines (SVM). An improvement toward the term weighting is conducted by using supervised term weighting with normalization, which we believe to be more suitable for the Indonesian pornographic text classification domain.

The result of experiment shows that our system—which implements SVM technique as classification technique and supervised term weighting with normalization as term weighting method—improves the accuracy and precision by 0.72% and 1.25%, respectively. Our system also reduces the level of over-blocking (non sensitive site is analyzed as a sensitive site) to 0.14% compared to the previous research.

**Keywords**—Sensitive Sites Detection and Filtering System, Pornography, Classification, Support Vector Machines, SVM, Supervised Term Weighting.

## INTISARI

Sistem pendeteksi dan penapis situs sensitif (pornografi) berbahasa Indonesia berdasarkan konten teks yang ada saat ini masih belum menghasilkan performa yang sesuai seperti yang diharapkan. Performa sistem ini, terutama pada tingkat akurasi, presisi, dan sensitivitas masih dapat ditingkatkan karena belum mencapai performa yang ideal yaitu 100%.

Penelitian ini berusaha untuk membangun sistem pendeteksi dan penapis situs sensitif yang memiliki performa yang lebih baik dari sistem yang dibangun sebelumnya. Performa dalam hal tingkat akurasi, presisi, dan sensitivitas ditingkatkan dengan mengaplikasikan teknik klasifikasi teks yang lebih sesuai, yaitu dengan menggunakan metode *Support Vector Machines* (SVM). Selain itu, penelitian ini juga melakukan perbaikan pada pembobotan fitur dengan menggunakan metode pembobotan fitur yang tergolong dalam metode *Supervised Term Weighting* dan normalisasi panjang dokumen.

Berdasarkan hasil yang didapatkan, sistem yang menggunakan model klasifikasi dengan teknik SVM dan metode pembobotan fitur *Supervised Term Weighting* dan normalisasi ini dapat meningkatkan performa sistem dalam hal tingkat akurasi dan presisi sebesar 0,72% dan 1,25% dari model klasifikasi sistem yang dibangun pada penelitian sebelumnya. Selain itu, sistem yang dibangun pada penelitian ini juga lebih baik dalam hal mengatasi permasalahan *over-blocking* (kesalahan sistem mendeteksi situs non-sensitif sebagai situs sensitif) dengan meningkatkan sensitivitas sistem sebesar 0,14% dari sistem yang dibangun pada penelitian sebelumnya.

**Kata Kunci:** Sistem Pendeteksi dan Penapis Situs Sensitif, Pornografi, Klasifikasi, *Support Vector Machines*, SVM, *Supervised Term Weighting*.