

ABSTRAK

Support Vector Machine (SVM) merupakan metode pembelajaran mesin *supervised learning* yang telah terbukti sebagai salah satu algoritma pembelajaran yang paling kuat untuk kategorisasi teks, tetapi karena pada prinsipnya bekerja secara *linear* sehingga untuk mengatasi masalah data *non-linear* dikembangkan fungsi kernel. Terdapat beberapa kernel *non-linear* salah satunya seperti kernel *polynomial*. Kernel *polynomial* memiliki parameter *degree* yang dapat disesuaikan sehingga mendapatkan hasil yang optimal. Penelitian ini melakukan analisis pengaruh kernel *polynomial* terhadap data *non-linear* yang diimplementasikan dalam identifikasi *hatespeech* berbahasa Indonesia pada komentar Instagram. Untuk mengetahui data yang dipakai berbentuk *linear* atau *non-linear*, maka digunakan teknik reduksi *Principal Component Analysis* (PCA) untuk memvisualisasikan data. Kernel yang digunakan yaitu kernel *polynomial*, *linear*, dan RBF.

Hasil pengujian menunjukkan bahwa metode *Support Vector Machine* menggunakan kernel *polynomial* memiliki performa yang lebih baik dibandingkan kernel *linear* dan kernel RBF dalam melakukan identifikasi data *non-linear* berjumlah 918 komentar yang terdiri dari 463 *non-hatespeech* dan 455 *hatespeech* dengan proses *split* data *training* dan *testing* 80:20, serta melalui tujuh proses *preprocessing* yaitu *case folding*, *cleaning*, *remove repetition character*, *tokenizing*, *normalisasi*, *stemming*, dan *stopword removal*. Selain itu, penentuan nilai parameter *C*, *degree*, dan *gamma* juga memiliki pengaruh dalam meningkatkan performa pada model. Performa terbaik kernel *polynomial* diperoleh dengan nilai parameter $C=0.1$, $degree=1$, dan $gamma=0.01$ yaitu akurasi 83.15%, presisi 83.45% dan *recall* sebesar 83.15%. Sedangkan kernel RBF diperoleh performa terbaik pada nilai parameter $C=2$ dan $gamma=0.001$ yaitu akurasi 80.43%, presisi 80.51%, *recall* 80.43%, dan kernel *linear* performa terbaik didapatkan dengan nilai parameter $C=0.01$ yaitu akurasi, presisi, dan *recall* sebesar 77.17%.

Kata kunci: analisis sentimen, ujaran kebencian, Instagram, data *non-linear*, kernel *polynomial*, *Support Vector Machine*, *Principal Component Analysis*

ABSTRACT

Support Vector Machine (SVM) is a supervised learning machine learning method that has proven to be one of the most powerful learning algorithms for text categorization, but because it works linearly in principle so to overcome the problem of non-linear data, a kernel function was developed. There are several non-linear kernels, one of which is the polynomial kernel. The polynomial kernel has degree parameters that can be adjusted to get optimal results. This research analyzes the effect of the polynomial kernel on non-linear data implemented in the identification of Indonesian-language hatespeech on Instagram comments. To find out the data used is linear or non-linear, the Principal Component Analysis (PCA) reduction technique is used to visualize the data. The kernels used are polynomial, linear, and RBF kernels.

The test results show that the Support Vector Machine method using the polynomial kernel has better performance than the linear kernel and RBF kernel in identifying non-linear data totaling 918 comments consisting of 463 non-hatespeech and 455 hatespeech with a split training and testing data process of 80:20, and through seven preprocessing processes namely case folding, cleaning, removing repetition characters, tokenizing, normalizing, stemming, and stopword removal. In addition, determining the value of the parameters C , degree, and gamma also has an influence in improving the performance of the model. The best performance of the polynomial kernel is obtained with a parameter value of $C = 0.1$, degree = 1, and gamma = 0.01, namely 83.15% accuracy, 83.45% precision and 83.15% recall. While the RBF kernel obtained the best performance at parameter value $C = 2$ and gamma = 0.001, namely accuracy 80.43%, precision 80.51%, recall 80.43%, and linear kernel the best performance is obtained with parameter value $C = 0.01$, namely accuracy, precision, and recall of 77.17%.

Keywords: sentiment analysis, hatespeech, Instagram, non-linear data, kernel polynomial, Support Vector Machine, Principal Component Analysis