ABSTRAK

Text mining yang merupakan perluasan *data mining* memiliki tujuan untuk mendapatkan informasi yang bermanfat dari data bertipe teks yang tidak terstruktur. *Twitter* merupakan media sosial berbasis teks dimana penggunanya dapat mengunggah ide atau pendapat secara bebas. Kebebasan untuk mengemukakan pendapat ini tidak lepas dampak negatif yang dihasilkan yaitu maraknya konten yang mengandung *hate speech*. Penelitian ini bertujan untuk mengetahui dan menganalisis performa algoritma *machine learning Support Vector Machine* (SVM) dalam melakukan klasifikasi *tweet* berbahasa Indonesia yang mengandung *hate speech*.

Dalam penelitian ini data yang digunakan terdiri dari 13.169 *tweet* dengan 5.561 data kategori *hate speech* dan 7.608 merupakan data dengan kategori non-*hate speech*. Pengujian dilakukan dengan mengeksplorasi seleksi fitur *information gain* dan *chi-square* dengan kombinasi *hyperparameter* pada algoritma SVM. Hasil pengujian menunjukkan bahwa nilai akurasi tertinggi dihasilkan dari kombinasi seleksi fitur *information gain*, *kernel* RBF, C = 10, dan gamma =1, dan k pada *k-fold cross validation* = 9 dengan nilai akurasi sebesar 86.68%.

Kata Kunci : *Text Mining,* Klasifikasi*, Hate Speech, Twitter, Support Vector Machine*

ABSTRACT

Text mining, which is an extension of data mining, has the aim of obtaining useful information from unstructured text type data. Twitter is a text-based social media where users can upload ideas or opinions freely. The freedom to express opinions cannot be separated from the resulting negative impact, namely the rise of content containing hate speech. This research aims to determine and analyze the performance of the Support Vector Machine (SVM) machine learning algorithm in classifying Indonesian language tweets that contain hate speech.

In this study, the data used consisted of 13.169 tweets, with 5.561 data in the hate speech category and 7608 data in the non-hate speech category. Testing was carried out by exploring information gain and chi-square feature selection with a combination of hyperparameters in the SVM algorithm. The test results show that the highest accuracy value is produced from a combination of information gain feature selection, RBF kernel, C =10, and gamma = 1 and k in k-fold cross validation = 9 with an accuracy result of 86.68%.

Keywords: Text Mining, Classification, Hate Speech, Twitter, Support Vector Machine.