ABSTRAK

Tingginya penggunaan perangkat seluler, mengakibatkan penyalahgunaan SMS sebagai media untuk melakukan kejahatan, seperti *spam* SMS, *phishing*, *malware*, dan lain-lain. Perlu adanya sebuah sistem yang bisa mengklasifikasi SMS dengan tujuan agar pengguna perangkat seluler tidak terganggu dengan *spam* SMS tersebut. Untuk bisa mengklasifikasi SMS, sistem tersebut memerlukan model *machine learning* yang memiliki akurasi tinggi agar bisa mengklasifikasi SMS *spam* dengan akurat.

Penelitian ini menggunakan *Support Vector Machine* (SVM) untuk mengklasifikasi teks SMS *spam* dengan *fastText* sebagai ekstraksi fitur. *Dataset* yang digunakan adalah *dataset* penelitian sebelumnya yang berisikan 1143 SMS yang terdiri dari 2 kelas, yaitu 569 *ham* dan 574 *spam.* Karena dataset SMS berisikan *string*, maka perlu dilakukan ekstraksi fitur terlebih dahulu sebelum digunakan untuk pembuatan model SVM. Untuk mengekstrak fitur dari dataset SMS tersebut, digunakanlah *fastText*. Kemudian setelah fitur dari dataset SMS terekstrak, maka langkah selanjutnya adalah membuat model *machine learning* untuk mengklasifikasi SMS *spam*. Model *machine learning* yang digunakan adalah *Support Vector Machine* atau SVM. Dari studi literatur, *fastText* merupakan ekstraksi fitur yang memperhitungkan struktur internal kata, posisi dan hubungan antar kata dan SVM merupakan metode yang lebih tahan mengalami *overfitting*. Metodologi penelitian ini dimulai dari pendahuluan, pengumpulan data, *preprocessing data*, ekstraksi fitur menggunakan *fastText*, pembuatan model SVM, pengujian model SVM, dan yang terakhir adalah hasil dan laporan.

Hasil pengujian model SVM menggunakan *K-Fold Cross Validation* dengan K sebesar 4 mendapatkan *accuracy* sebesar 95,97%, *precision* sebesar 94,27%, *recall* sebesar 97,88%, dan *f1-score* sebesar 96,03% dengan parameter C sebesar 1, dan *gamma* sebesar 0,01.

Kata Kunci: SMS *spam*, *Support Vector Machine*, *fastText*, klasifikasi teks

ABSTRACT

The high use of mobile devices has resulted in the misuse of SMS as a medium to commit crimes, such as SMS spam, phishing, malware, and others. There needs to be a system that can classify SMS with the aim that mobile device users are not disturbed by the SMS spam. To be able to classify SMS, the system requires a machine learning model that has high accuracy in order to accurately classify SMS spam.

This research uses Support Vector Machine (SVM) to classify SMS spam text with fastText as feature extraction. The dataset used is the previous research dataset which contains 1143 SMS which consists of 2 classes, namely 569 ham and 574 spam. Because the SMS dataset contains strings, it is necessary to extract features first before being used for SVM modeling. To extract features from the SMS dataset, fastText is used. Then after the features of the SMS dataset are extracted, the next step is to create a machine learning model to classify SMS spam. The machine learning model used is the Support Vector Machine or SVM. From the literature study, fastText is a feature extraction that takes into account the internal structure of words, positions and relationships between words and SVM is a method that is more resistant to overfitting. The research methodology starts from introduction, data collection, data preprocessing, feature extraction using fastText, SVM model creation, SVM model testing, and the last is results and reports.

The results of the SVM model test using K-Fold Cross Validation with K of 4 getting an accuracy of 95.97%, precision of 94.27%, recall of 97.88%, and f1-score of 96.03% with parameter C of 1 , and a gamma of 0.01.

Keyword: *SMS spam, support vector machine, fastText, text classification*