

INTISARI

Klasifikasi File Berisi *Secret Key* Menggunakan *Character Level* CNN dan Shannon Entropy

Oleh

Rafie Muhammad

20/466429/PPA/05995

Perkembangan teknologi saat ini berbanding lurus dengan banyaknya proyek teknologi informasi yang dibuat. Proyek ini biasanya berupa suatu program yang dibentuk dari suatu kode. Untuk memudahkan kontrol versi dan kontribusi antar *developer*, biasanya digunakan *tools* git yang ditempatkan pada suatu *online platform* seperti Github. Jika suatu *repository* di suatu *online platform* mengandung informasi sensitif seperti variabel *secret key* serta *repository* tersebut dapat diakses oleh khalayak umum, maka hal tersebut dapat dimanfaatkan untuk membahayakan perusahaan atau organisasi yang memiliki *repository* tersebut. Beberapa cara telah dilakukan sebelumnya untuk mendeteksi variabel *secret key* ini dengan menggunakan metode regex serta menggunakan *machine learning* untuk meminimalisasi hasil *false positive*. Dalam penelitian ini, akan coba dibangun suatu model klasifikasi teks atau file yang berisi variabel *secret key* dengan menggunakan *character level* CNN dan *shannon entropy*. Algoritma *shannon entropy* digunakan pada saat *pre-processing* untuk mengambil kelompok kata dengan nilai *entropy* terbesar dan *character level* CNN akan digunakan karena dinilai cocok untuk kasus ini karena nilai *secret key* yang abstrak serta tidak memerlukan konteks sekuens untuk kata *secret key* tersebut pada suatu file. Hasil dari penelitian ini dapat memperoleh nilai akurasi sebesar 96.1 % untuk nilai akurasi pada klasifikasi file yang berisi *secret key*.

Kata kunci : *character level* CNN, *shannon entropy*, *secret key*, teks klasifikasi, *git repository*, *deep learning*

ABSTRACT

FILE CLASSIFICATION CONTAINING SECRET KEY USING CHARACTER LEVEL CNN AND SHANNON ENTROPY

By

Rafie Muhammad

20/466429/PPA/05995

Current technological developments are directly proportional to the number of information technology projects made. This project is usually in the form of a program formed from a code. To facilitate version control and contributions between developers, git tools are usually used which are placed on an online platform such as Github. If a repository on an online platform contains sensitive information such as secret key variables and the repository can be accessed by the general public, then this can be used to harm the company or organization that owns the repository. Several ways have been done before to detect this secret key variable by using the regex method and using machine learning to minimize false positive results. In this research, we will try to build a text or file classification model that contains secret key variables using the character level CNN and shannon entropy. The Shannon entropy algorithm is used during pre-processing to retrieve the group of words with the largest entropy value and the CNN character level will be used because it is considered suitable for this case since the value of the secret key is abstract and does not require a sequence context for the said secret key in a file. The results of this study were able to obtain an accuracy value of 96.1 % for the accuracy value for the classification of files containing *secret key*.

Keywords : *character level CNN, shannon entropy, secret key, text classification, git repository, deep learning*