INTISARI

Praktik *data mining* semakin banyak dilakukan di era yang semakin modern ini. Hal ini disebabkan oleh kemudahan dalam mendapatkan data yang didukung oleh semakin banyaknya sumber data yang berasal dari sumber yang terbuka. Salah satu sumber data terbuka tersebut adalah situs web pelaporan yang dikelola oleh Pemerintah Provinsi Jawa Tengah, yaitu LaporGub. Masyarakat yang memberikan laporan seringkali tidak sadar bahwa mereka memberikan informasi sensitif yang terlalu banyak sehingga dapat membahayakan diri mereka sendiri. Di sisi lain, banyak masyarakat yang mempersoalkan mengenai keamanan dan privasi mereka ketika memberikan laporan melalui situs web LaporGub. Untuk mengatasi masalah tersebut, pengelola situs web dengan data terbuka dapat menerapkan model pembelajaran mesin *Named Entity Recognition* (NER) yang dibuat dalam *framework* situs web mereka untuk menjaga *personally identifiable information* (PII) atau informasi sensitif dari penggunanya. Model NER yang dibuat menggunakan pustaka *spaCy* dan berfokus dalam mendeteksi entitas sensitif pada teks bahasa Indonesia. Metodologi yang digunakan meliputi pengumpulan data dengan *web scraping*, verifikasi data, *preprocessing* data, pelabelan data, pembuatan model NER, dan evaluasi kinerja model NER. Hasil penelitian menunjukkan bahwa seluruh model NER yang dibuat dapat mendeteksi *personally identifiable information* dengan hasil dan performa yang baik. Hasil tersebut meliputi 6 model *deep learning* non-*Transformer* dengan akurasi *f1-score* keseluruhan di antara 87-90% dan model terbaik merupakan model *deep learning* berbasis *Transformer* menggunakan model *pre-trained* dari IndoBERT dengan akurasi *f1-score* sebesar 91%. Kesimpulan dari penelitian ini adalah model NER yang dibuat dapat meningkatkan privasi dan tingkat keamanaan pada situs web dengan data terbuka untuk melindungi privasi dan data pengguna.

Kata kunci : *Named Entity Recognition*, Keamanan, Privasi, *Personally Identifiable Information*, SpaCy

ABSTRACT

Data mining practices are increasingly being carried out in this more modern era. This is due to the ease of obtaining data and supported by the increasing number of data sources originating from open sources. One such open data source is a reporting website managed by the Central Javal Provincial Government, named LaporGub. Communities who use this website are often not aware that they are providing too much sensitive information that can put them in dangerous condition. On the other hand, many people are concerned about their security and privacy when submitting reports via LaporGub website. To solve this problem, open data website managers can implement Named Entity Recognition (NER) machine learning models built into their website frameworks to safeguard Personally Identifiable Information (PII) of their users. The NER model is built using the spaCy library and focuses on detecting PII in Indonesian text. The methodology used includes data collection by web scraping, data verification, data preprocessing, data labeling, NER modeling, and NER model performance evaluation. The results showed that all NER models created could detect PII object with good results and performances. These results include six non-Transformer deep learning model with overall f1-score accuracy between 87-90% and the best model is achieved by Transformer-based deep learning model using pre-trained model from IndoBERT with an overall f1-score accuracy of 91%. The conclusion from this research is that all NER models built in this research can increase privacy and security levels on open data website to protect privacy and user data.

Keywords : *Named Entity Recognition*, *Privacy*, *Personally Identifiable Information*, *Security*, SpaCy