

Rancang Bangun Fitur Data Loss Prevention untuk Mendeteksi Data Pribadi Saat Transit Berdasarkan Undang-Undang Pelindungan Data Pribadi Menggunakan Zeek, Apache Tika, dan Model Fine-Tuned BERT = Design and Construction of Data Loss Prevention Feature to Detect Personal Data in Transit Based on Personal Data Protection Act Using Zeek, Apache Tika, and Model Fine-Tuned BERT

Satya Ananda Sulistio, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920571296&lokasi=lokal>

Abstrak

UU Pelindungan Data Pribadi (PDP) mewajibkan pengendali data pribadi untuk menggunakan sistem elektronik yang aman dan bertanggung jawab supaya terhindar dari kehilangan data, sebagaimana diatur dalam Pasal 39. Untuk memenuhi tuntutan tersebut, penelitian ini merancang salah satu fitur data loss prevention (DLP), yaitu pendeteksian data pribadi, yakni NIK, nomor kartu debit/kredit, dan nomor BPJS, ketika file berisi informasi sensitif tersebut sedang bergerak di jaringan (data in transit). Ketiga sampel ini mewakili antarmuka utama dalam pengaksesan data pribadi spesifik warga negara Indonesia, seperti diklasifikasikan oleh Bab 2 UU PDP. Penelitian ini mengajukan suatu pipeline pendeteksi yang dibangun menggunakan teknologi gratis dan sumber terbuka, memanfaatkan kemampuan OCR dari Apache Tika serta model IndoBERT yang telah di-fine-tune dengan data sintesis hasil pembangkitan LLM Llama 3.1 8B. Hasil evaluasi menunjukkan skor F1 rata-rata sebesar 89.75%. Pengujian skalabilitas mengungkapkan hambatan pada pemrosesan file berukuran besar akibat mekanisme multi-threading dari Apache Tika, sedangkan pengujian latensi menegaskan bahwa inferensi model BERT dapat berjalan cepat meski hanya menggunakan CPU dan lebih cepat dibandingkan teknik pendahulu, seperti teknik regex. Selain itu, penerapan pemprofilan pengguna yang didukung dalam alur pipa sistem ini juga selaras dengan prinsip Zero Trust Access.

.....Private Data Protection (PDP) Law enforces private data controller to utilize a secure and reliable electronic system in order to prevent data loss, as mandated by Article 39. To answer that challenge, this research develops one of data loss prevention (DLP) features, which is private data detection, such as NIK, debit/credit card number, and BPJS number, when files containing them are being transmitted over the network (data in transit). The three data samples are gateways to accessing multiple specific private data as classified by Chapter 2 of PDP Law. The proposed detection pipeline is built entirely with free, open-source technologies, leveraging Apache Tika's OCR capabilities and an IndoBERT model fine-tuned on synthetic data generated by the LLM Llama 3.1 8B. Evaluation results show an average F1 score of 89.75%.

Scalability tests reveal a bottleneck when processing large files due to Apache Tika's multi-threading limitations, while latency tests confirm that the fine-tuned BERT inference runs quickly on CPU alone and offers speed-up compared to a classical technique, regex. Additionally, the implementation of user profiling aligns with the principles of Zero Trust Access.