Universitas Indonesia Library >> UI - Skripsi Membership

Pengembangan Sistem Prediksi Situs Web Berbahaya Menggunakan Large Language Model Qwen3 dengan Knowledge Distillation dari DeepSeek-R1 = Development of a Malicious Website Prediction System Using Qwen3 Large Language Model with Knowledge Distillation from DeepSeek-R1

Rizki Awanta Jordhie, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920571267&lokasi=lokal

---

Abstrak

Penelitian ini bertujuan untuk mengembangkan sistem prediksi situs web berbahaya menggunakan pendekatan Large Language Model (LLM) Qwen3 yang di fine-tuning melalui proses knowledge distillation dari model DeepSeek-R1 sebagai model guru. Sistem ini dirancang untuk mengklasifikasikan domain dan konten situs web ke dalam empat kategori: jinak (benign), perjudian (gambling), pornografi (pornography), dan berbahaya (harmful). Dataset yang digunakan berasal dari Common Crawl dan sumber relevan lain, telah dikurasi dan dilabeli secara otomatis dengan reasoning hasil distilasi DeepSeek-R1 untuk mencerminkan skenario dunia nyata. Evaluasi sistem dilakukan menggunakan metrik presisi, recall, F1-score, dan akurasi. Hasil eksperimen menunjukkan bahwa model Qwen3-4B yang di fine-tuning dengan knowledge distillation mampu mencapai akurasi hingga 96,25% dan F1-Makro 0,8378, menandakan peningkatan signifikan dibandingkan baseline. Penelitian ini membuktikan efektivitas transfer pengetahuan dari DeepSeek-R1 ke Qwen3 dalam meningkatkan performa klasifikasi situs web berbahaya, serta memberikan kontribusi pada pengembangan sistem deteksi berbasis LLM yang lebih adaptif dan transparan.
.....This study aims to develop a malicious website prediction system using Qwen3-based Large Language Model (LLM) fine-tuned through a knowledge distillation process from DeepSeek-R1 model as the teacher. The system is designed to classify website domains and content into four categories: benign, gambling, pornography, and harmful. The dataset used is sourced from Common Crawl and other relevant sources, curated and automatically labeled with reasoning distilled from DeepSeek-R1 to reflect real-world scenarios. System evaluation was conducted using precision, recall, F1-score, and accuracy metrics. Experimental results show that the Qwen3-4B model fine-tuned with knowledge distillation achieved up to 96.25% accuracy and a macro F1-score of 0.8378, indicating a significant improvement over the baseline. This study demonstrates the effectiveness of knowledge transfer from DeepSeek-R1 to Qwen3 in improving malicious website classification performance and contributes to the development of more adaptive and transparent LLM-based detection systems.