Paper ID: 1570940847

Paper Title: De-identification of Thai Free-text Clinical Notes

Authors: Kerdkiat Suvirat, Sawrawit Chairat, Kanakorn Horsiritham, Chanon Kongkamol and Sitthichok Chaichulee (Prince of Songkla University, Thailand)

Email: 6210210229@psu.ac.th

Abstract

The use of electronic health record (EHR) systems enables the storage of extensive historical patient data. However, accessing this data for secondary use is limited due to the presence of personally identifiable information (PII) and protected health information (PHI), which raises ethical concerns and requiring compliance with data protection regulations. In the United States, Health Insurance Portability and Accountability Act (HIPAA) governs PHI use, mandating strict security and privacy measures. De-identification research must follow HIPAA guidelines to safeguard patient confidentiality. Similarly, in Thailand, the Personal Data Protection Act (PDPA) regulates personal data, including health information. Complying with the PDPA is crucial for research on Thai patient data, ensuring adherence to data de-identification provisions. This study investigated the use of transformer-based language models for the de-identification of Thai clinical notes. We used a dataset comprising 6,134 clinical notes, with PHI entities annotated by experts. We explored three pretrained language models: Multilingual BERT, WangChanBERTa, and MEDPSU-RoBERTa, for de-identification purposes. The results demonstrate that the model based on MEDPSU-RoBERTa outperforms the other models in overall de-identification performance, emphasizing the significance of domain-specific training data. It achieved a precision of 0.9320, a recall of 0.9718, and a F1 score of 0.9552. Our algorithm was able to identify and remove six PHI categories: age, contact, datetime, identifier, location, and name. The study also confirms the potential of language models for effective de-identification and highlights the benefits of domain-specific training for anonymizing clinical text.