Deidentification and Generation of Indian Medical Text

Sanjeet Singh, *IIT Kanpur*

***Abstract*—In the era of artificial intelligence, high-quality data is essential to revolutionizing innovation in domains such as medicine, finance, education, etc. However, in the medical field, privacy concerns make hospitals hesitant to share patient data, which limits access to real-world information. Although a few public datasets exist, most come from Western countries and are unsuitable for building viable solutions for Indian hospitals or patients. Medical data is highly dependent on demographics and geographic context, so data from other regions cannot directly support India-specific applications. So, in this paper, to mitigate privacy concerns, we build a robust de-identification technique to de-identify Indian medical texts. We need a lot of real patient data to build the deidentification tool. We do not have access to that much data, so we utilize the capability of LLMs to generate synthetic data.**