Stage M2: Vérification de la confidentialité de corpus cliniques synthétiques

Contexte

Les travaux du stage s'inscrivent dans le cadre du projet ANR CoDeinE (artificial text COrpus DEsIgNed Ethically) qui fait l'objet d'une collaboration entre le Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) et le CEA.

Le projet s'intéresse à la confidentialité des données sensibles en domaine de spécialité, avec une application au domaine biomédical. Dans le domaine clinique, l'accès à des textes, même désidentifiés, est restreint pour des raisons de confidentialité. Ainsi, la problématique de la confidentialité et de la sécurité des modèles se pose pour les modèles de langue appris sur des textes sensibles et pour des corpus de textes synthétiques générés avec ces modèles. Le projet cherche à explorer, à l'aide de méthodes issues de la sécurité des données, dans quelle mesure des informations confidentielles issues des corpus sensibles utilisés pour l'entrainement de modèles de langue peuvent être extraites des modèles eux-mêmes et de textes synthétiques produits par ces modèles. Nous nous plaçons dans un cadre expérimental dans lequel les corpus d'entrainement sensibles font l'objet d'une désidentification. Des travaux précédents (Vakili et al. 2022) ont montré la robustesse des modèles aux attaques sur les données identifiantes. Cependant, la combinaisons d'élément phénotypiques (par exemple, age et pathologies d'un patient) peut également être identifiante. Le projet cherchera à explorer cette dimension : le recouvrement de triplets d'éléments cliniques, selon que l'attaquant dispose d'un modèle ou d'un corpus synthétique.

Approche envisagée pour le stage

Étant donné un texte clinique synthétique généré à l'aide d'un modèle de langue, on cherche à savoir si des informations confidentielles concernant les patients dont les données sont présentent dans les corpus ayant servis à construire le modèle peuvent se retrouver dans les textes générés.

Les travaux précédents ont porté sur la présence des données directement identifiantes (nom, prénom, adresse ...). Dans ce stage, nous nous intéressons à informations phénotypiques ç-à-d des combinaisons uniques d'informations cliniques permettant d'identifier une personne (femme 30 ans, cancer du sein, 5 enfants, appendicite ...).

On va par exemple s'intéresser à extraire des triplets ou des éléments qui peuvent correspondre à des vrai patients dans les données générées.

Il y a deux cas à prévoir :

— On dispose d'un corpus synthétique et d'un corpus d'apprentissage
— On dispose d'un modèle de génération et d'un corpus d'apprentissage

Les expériences seront réalisés avec des corpus de cas clinique en Français et des documents cliniques issus d'un groupement hospitalier partenaire du projet.

La littérature abonde d'approches permettant les attaques d'extraction de données privées de modèles d'apprentissage profond. Læ candidat·e pourra explorer les attaques d'extraction de données privées sur les gros modèles de langues et leur praticabilité[Carlini et al., 2021], ainsi que les attaques par appartenance (*membership inference attacks* [Carlini et al., 2022, Mahloujifar et al., 2021]).

Profil recherché

Vous suivez une formation d'ingénieur ou de master spécialité machine learning / linguistique informatique / data science et vous disposez :

- de connaissances préalables dans les domaines du traitement automatique des langues, de l'apprentissage automatique et de l'apprentissage profond ;
- d'un intérêt prononcé pour la recherche et les nouvelles technologies et êtes à l'aise en anglais ;
- de créativité et de motivation ;
- de solides compétences en programmation et maitrisez l'un des frameworks DL (TensorFlow ou PyTorch).

Stage de à pourvoir sur Orsay au sein du LISN (91), indemnisation forfaitaire.
Contact : `sahar.ghannay@lisn.upsaclay.fr` `Julien.GIRARD2@cea.fr`
Ce stage pourra déboucher sur une thèse au sein du laboratoire.

Références

[Carlini et al., 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. (2022). Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

[Carlini et al., 2021] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium (USENIX Security 21)*.

[Mahloujifar et al., 2021] Mahloujifar, S., Inan, H. A., Chase, M., Ghosh, E., and Hasegawa, M. (2021). Membership Inference on Word Embedding and Beyond. *arXiv :2106.11384 [cs]*.