Stage M2 – 2023 (durée 5-6 mois)

Vers un modèle explicable pour la détection d'infox sur des données médicales basée sur des méthodes d'apprentissage profond

Equipes SDC (Science des Données et Connaissances)
et CSTB (Systèmes Complexes et Bioinformatique Translationnelle)
Laboratoire ICube de Strasbourg

Contexte

Les systèmes autonomes intelligents dits à "boîte noire" reposent sur des algorithmes d'apprentissage basés sur des réseaux de neurones profonds (« deep learning »), dont la structure complexe masque le fonctionnement interne. L'exigence d'une meilleure compréhension des décisions des algorithmes d'apprentissage profond devient un enjeu sociétal, et requiert le développement de techniques permettant de comprendre leur fonctionnement ou d'expliquer leurs décisions. Ce stage s'inscrit dans le cadre du projet DEEPISH (Deep lEarning ExPlainabilIty through Symbolic approacHes), qui a pour objectif de proposer un modèle reposant sur des techniques de raisonnement symbolique - graphes de connaissances, ontologies ou règles - permettant d'expliquer les décisions de systèmes basés sur un apprentissage profond.

Problématique

Le modèle d'explicabilité développé est lié à la nature des données du domaine d'application et aux types d'algorithmes d'apprentissage profond utilisés. Nous nous intéressons dans ce travail à des données textuelles issues du domaine médical. L'objectif est de proposer une méthode de détection d'informations fallacieuses ou *infox* ("fake news") issues des réseaux sociaux ou plus généralement de données collectées sur internet. Cette détection devra s'accompagner d'un schéma d'explication.

Travail demandé

La détection d'infox est une tâche de classification de textes qui consiste à prédire si un texte court est vrai ou faux (classification binaire). Les exemples ci-dessous sont issus de (Ayoub et al., 2021) :

- Faux : "COVID-19 only affects the old."
- Vrai : "Being able to hold your breath for 10 s or more without coughing or feeling discomfort DOES NOT mean you are free from COVID-19."

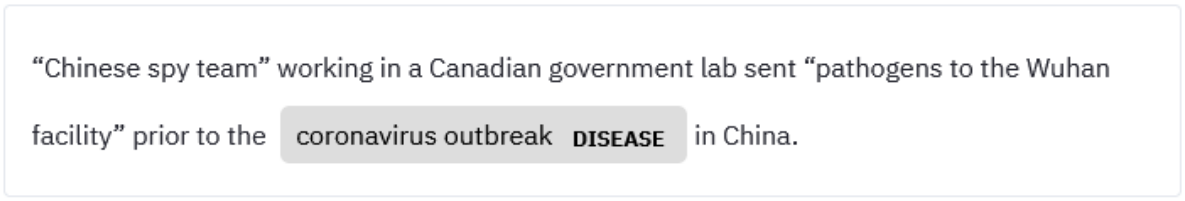
Les méthodes les plus récentes de classification de textes reposent sur des modèles de langue pré-entraînés à l'aide de grandes quantités de données textuelles, du type "BERT". Les

modèles de langues sont ensuite affinés pour une tâche précise (classification de texte par exemple), selon un processus d'"affinage" (*fine tuning*) qui utilise des données d'entraînement annotées.
Dans les tâches de classification de textes, les décisions prises par les algorithmes d'apprentissage sont expliquées par l'importance de certaines caractéristiques, généralement les mots (Ribeiro et al., 2016). Or les infox sur les données médicales utilisent des termes renvoyant à des entités et concepts médicaux qu'il est possible d'annoter automatiquement à l'aide d'outils comme HunFLAIR (Weber et al., 2021) ou Scispacy[1]. La figure ci-dessous montre quelques résultats d'annotation avec Scispacy :

Entities are linked to the Unified Medical Language System (UMLS).

	text	Canonical Name	Concept ID	TUI(s)	Score	start	end
0	pathogens	Pathogenic organism	C0450254	T001	0.858097	13	14
1	outbreak	Disease Outbreaks	C0012652	T067	1	23	24
2	China	China	C0008115	T083	1	25	26

Specialized NER



	text	label_	start	end	start_char	end_char
0	coronavirus outbreak	DISEASE	22	24	108	128

Il s'agira donc d'exploiter ces concepts et entités pour améliorer l'explicabilité des modèles de détection d'infox sur des données médicales. Les concepts extraits pourront s'organiser au sein d'un graphe de connaissances ou d'une ontologie. La détection pourra utiliser un modèle

[1] https://allenai.github.io/scispacy/

d'apprentissage profond de type « Transformer », basé sur le mécanisme d'attention (Stahlberg, 2020).

Perspectives

Un stage de Master effectué en 2022 dans le cadre du projet DEEPISH (Wehrli, 2022), a conduit au choix du modèle d'explicabilité « TREPAN Reloaded » (Besold & al, 2020), qui permet d'expliquer les décisions de réseaux de neurones quelconques codés sur Tensorflow / Keras. Ce système travaille sur des données textuelles, en s'appuyant sur des ontologies du domaine considéré, afin de produire des arbres de décision compréhensibles par un humain.

Après avoir développé un prototype de modèle d'apprentissage profond pour la détection des *infox* sur des données médicales s'appuyant sur un graphe de connaissances ou une ontologie, on pourra étudier la possibilité d'interfacer ce modèle avec « TREPAN Reloaded », afin de produire un modèle d'explicabilité plus général.

Environnement de travail et encadrement

Le stage s'effectuera dans les locaux du laboratoire ICube à Illkirch, au sein de l'équipe SDC Science des Données et Connaissances. L'encadrement sera assuré par :

- Delphine Bernhard (traitement automatique des langues, LiLPa - Faculté des langues), dbernhard@unistra.fr ,
- Anne Jeannin-Girardon (apprentissage profond, CSTB - ICube), jeanningirardon@unistra.fr ,
- Stella Marc-Zwecker (modélisation de connaissances, SDC - ICube), stella@unistra.fr

Bibliographie

Ayoub J., Yang X.J., and Zhou F. (2021). *Combat COVID-19 infodemic using explainable natural language processing models*, Information Processing & Management, 58, 4, p.p. 102569.

Besold T.R., Confalonieri R., T. Weyde T. and F.M. del Prado Martín (2020), *Trepan reloaded : A knowledge-driven approach to explaining artificial neural networks.* Technical report, ECAI, 2020.

Ribeiro M.T., Singh S., et Guestrin C. (2016). *« Why Should I Trust You? »: Explaining the Predictions of Any Classifier,* Consultable à http://arxiv.org/abs/1602.04938 [Accédé le 18 août 2016].

Stahlberg, F. (2020). *Neural machine translation: A review.* Journal of Artificial Intelligence Research, 69, 343-418.

Weber L., Sänger M., Münchmeyer J., Habibi M., Leser U., et Akbik A. (2021). *HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition*, Bioinformatics, 37, 17, p.p. 2792□2794.

Wehrli L. (2022). *Explicabilité des réseaux profonds au moyen d'approches symboliques*, Mémoire de stage de Master d'informatique, ICube, 2022