Encodeur hybride pour la détection automatique de désinformation

Géraud Faye*, Sylvain Gatepaille*, Guillaume Gadek*, Souhir Gahbiche*

* Airbus, Élancourt, France
{geraud.faye, sylvain.gatepaille, guillaume.gadek, souhir.gahbiche}@airbus.com

Résumé. L'encodage de texte pour des tâches de classification repose aujourd'hui grandement sur de larges modèles de langage difficilement explicables et nécessitant de grandes quantités de données pour fonctionner. Ces modèles sont à la base de tâches de classification comme la détection de désinformation, importante aujourd'hui. Récemment, les approches hybrides entre l'apprentissage profond et l'IA symbolique tentent de surpasser les performances des modèles à base d'attention en introduisant du raisonnement dans le processus de décision pour le rendre moins opaque à l'utilisateur. Dans cet article, nous proposons CATS, un mécanisme d'attention basé sur la compréhension sémantique des documents, améliorant les performances des modèles neuronaux équivalents, réduisant le besoin en données annotées et facilitant l'explicabilité de la décision.

1 Introduction

Avec le récent et fort développement des réseaux sociaux, la diffusion de désinformation est devenue de plus en plus présente, au point où une majorité la considère comme une menace pour la démocratie [1]. Les réseaux sociaux deviennent l'unique source d'information pour de plus en plus de personnes [2], ce qui en fait le terrain idéal pour la désinformation. Il s'agit d'une certaine forme de mésinformation (information de mauvaise qualité) s'appuyant sur des biais cognitifs (Greifeneder et al., 2020) afin d'influer sur l'opinion publique.

Étant par nature liée à l'actualité et à la politique, la désinformation peut avoir un impact important lorsqu'elle est utilisée pour manipuler les votes lors d'élections majeures (élections américaines de 2016 ou référendum du Brexit) ou pour impacter la santé publique (récente crise du Covid-19 ou les vaccins en général). La manière dont ces informations sont rédigées est idéale pour les réseaux sociaux, car elle fait réagir et incite au partage. De plus, il est très difficile de réparer les dommages causés une fois qu'elles ont été beaucoup lues (loi de Brandolini). Cela rend important leur détection avant qu'elles ne soient largement partagées. La grande quantité d'information postée quotidiennement rend l'automatisation de cette tâche de détection cruciale.

Les travaux actuels se concentrent soit sur des facteurs de style des articles (utilisation de *features* de modèles transformers, règles symboliques), soit sur des facteurs de propagation

1. www.civica.eu/fake-news-and-democracy/
2. en.wikipedia.org/wiki/Social_media_as_a_news_source