VOISINS NACHBARN
L'information libre et mitoyenne

LORRAINE – SARRE – RHÉNANIE-PALATINAT

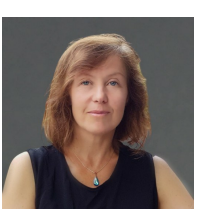
M-Phasis débusque la haine sur internet

Le projet franco-allemand M-Phasis, conduit par des chercheurs en sciences humaines et en informatique des universités de Lorraine, de Sarre et de Mayence, fourbit des armes innovantes pour juguler la haine en ligne.

DR

Romain Gascon *mercredi 18 octobre 2023*

Depuis les cours des écoles jusqu'aux soubresauts géopolitiques du moment, la haine en ligne semble pouvoir balayer en relative impunité toutes les parcelles de nos existences. Des chercheurs en sciences humaines et en informatique de plusieurs composantes des universités de Lorraine, de Sarre et de Mayence ont entrepris de disséquer des messages haineux pour nourrir la riposte avec le projet M-Phasis, lancé en 2018.



Irina Illina, maîtresse de conférences à l'IUT Nancy Charlemagne et chercheuse au Multispeech du Loria.

« Il n'y avait pas vraiment de définition unifiée de la haine. Cette notion n'avait pas été abordée dans la littérature scientifique », note Irina Illina, maîtresse de conférences à l'IUT Nancy Charlemagne et chercheuse au Multispeech du Loria, à l'origine du projet avec Angeliki Monnier, directrice du Centre de recherche sur les médiations et professeure en Sciences de l'information et de la communication à l'Université de Lorraine.

Après avoir défini scientifiquement le concept, les chercheurs ont collecté des commentaires en français et en allemand sur les sites internet de médias d'information des deux pays et sur le réseau social Twitter (devenu X, dans l'intervalle). Le corpus établi a été soigneusement étudié et annoté pour faire émerger un protocole solide et global de détection des discours de haine. Aux chercheurs en informatique est revenue la tâche de développer un outil pour l'automatiser, baptisé Human (Hierarchical universal modular annotator).

L'implicite n'a qu'à bien se tenir

« Notre logiciel évalue la probabilité de la présence de la haine dans un message textuel », résume Irina Illina. Il permet de gagner en efficacité dans l'analyse des flux de messages, à la fois en termes de volume et de rapidité, mais aussi dans le degré de finesse. A la détection de la haine explicite s'ajoute celle de la haine implicite. Parmi plusieurs travaux de thèse irrigués par le projet, celle qui sera soutenue en novembre prochain s'intéresse par exemple aux expressions dites « polylexicales », dont le sens global ne peut être déduit de la combinaison des sens des mots qui la composent.

Intelligence artificielle et citoyenne

Les résultats de M-Phasis présentent un intérêt manifeste pour les médias en ligne et les réseaux sociaux. La législation de l'Union européenne impose aux gestionnaires de plateformes de supprimer les messages à caractère haineux dans les 24 heures qui suivent leur publication. *« Cette détection est très coûteuse »*, note Irina Illina. Conduit et financé par des institutions publiques (Agence nationale de la recherche française et son homologue allemande Deutsche Forschungsgemeinschaft), M-Phasis s'inscrit dans le cadre du projet Olki (Open language and knowledge for citizens) développé par Lorraine Université d'excellence, qui promeut le développement d'une intelligence artificielle transparente. Scientifiques ou citoyens, tout le monde peut se saisir des résultats de M-Phasis, disponibles en open source.

Chat GPT générateur de haine

A ce stade, il est difficile de cerner l'impact que M-Phasis peut déjà avoir sur la haine en ligne. A tout le moins, il suscite l'intérêt. Il aura notamment les honneurs d'une conférence au Japon en novembre. Mais les enseignements sont d'ores et déjà nombreux. *« Nous nous sommes rendus compte que la haine est souvent propagée par quelques personnes seulement. Cela peut être utile pour la détecter. (…) Quant aux LLM* [Large language models, type Chat GPT, NDLR]*, ils sont efficaces pour mieux la tracer. Mais ils sont eux-mêmes capables de générer des messages haineux. Comment « débiaiser » ces modèles ? »*, interroge Irina Illina. A l'aune des résultats et enseignements de M-Phasis, clos à l'été 2022, les chercheurs espèrent obtenir de nouveaux financements pour pouvoir mieux étouffer la haine en ligne dans l'œuf et dans les textes, mais aussi dans les sons et les images.