Proposition de thèse – Engagement et émotions des utilisateurs dans la caractérisation de contenus toxiques sur Internet

Co-direction : Valentina DRAGOS (ONERA, Palaiseau)/ Delphine BATTISTELLI (MoDyCo, Nanterre)
Salaire : 2488€/brut mensuel
Début : septembre 2025
Nationalité européenne requise
Date limite de candidature : 15 avril 2025

A. Problématique générale et contexte

Le sujet de thèse aura pour objectif central l'étude de l'apport de traits linguistiques de nature sémantique dans la caractérisation et la détection de contenus en ligne de nature toxique (e.g. contenu haineux, contenu extrémiste, contenu discriminant). Plus précisément, les connaissances linguistiques examinées seront en lien : (i) d'une part avec l'analyse du degré d'engagement (e.g. « X est Y » vs. « je vous jure/j'ai la preuve que X est Y ») ; (ii) d'autre part avec l'analyse des émotions (e.g. « quel horrible personnage ! »).

Concernant le trait linguistique (i), la problématique de la thèse prend place au sein de travaux visant à examiner de quelle manière les émotions peuvent constituer un trait linguistique pertinent en vue de l'amélioration de systèmes d'identification automatique de discours haineux en ligne, visée encore largement inexplorée en traitement automatique des langues (TAL) [1] et ce alors que plusieurs études ont montré pourtant que la présence d'émotions - en particulier négatives - est un facteur favorable à la propagation de contenus sur Internet (e.g.[2], [3]). Concernant le trait linguistique (ii), la thèse prend place au sein de réflexions menées autour du fait que le degré d'engagement constitue un signe de radicalité et l'on sait que des positionnements radicaux sont de nature à être à plus diffusés que d'autres sur internet, contribuant potentiellement à un phénomène de désinformation [4].
Le sujet prend place dans le contexte sociétal actuel dans lequel sont perceptibles diverses menaces liées à l'adhésion de différentes catégories de personnes aux à des idéologies et mouvances extrémistes (e.g. extrême droite, masculinisme). L'utilisation des plateformes sociales, adoptées par un nombre croissant d'utilisateurs, constitue un vecteur important de diffusion de ces idéologies et la production massive de données disponibles en ligne a déclenché une importante activité de recherche autour de l'analyse de la subjectivité, entendue comme l'expression des goûts, opinions, croyances ou convictions des utilisateurs [5].

B. Objectifs scientifiques

Les objectifs scientifiques s'organiseront autour de deux volets principaux : 1. La construction de ressources linguistiques sous la forme de corpus annotés selon les traits (i) et (ii) ainsi que la construction d'ontologies relevant globalement de la représentation de la subjectivité ;
2. Le développement de méthodes de détection automatique de contenus toxiques capables d'exploiter les connaissances modélisées par ces ressources.

1. Construction de ressources linguistiques

Plusieurs corpus représentatifs seront constitués pour décrire les contenus toxiques sur

Internet. La construction du corpus s'appuiera sur une liste d'éléments (mots clés, patrons lexicaux) dont l'équipe d'accueil dispose actuellement grâce aux résultats de projets antérieurs, dont notamment FLYER (Artificial intelligence for extremist content analysis, projet ANR ASTRID 2019). Cette liste sera enrichie et mise à jour en explorant de nouvelles données collectées au cours de la thèse.
L'annotation des données sera réalisée en utilisant des schémas d'annotation existants, pour décrire des unités linguistiques émotionnelles [6] et pour décrire le degré d'engagement [7], [8].

On visera également ici à construire des ontologies décrivant la subjectivité, incluant les traits d'émotions et d''engagement des utilisateurs, et s'appuyant là aussi sur des travaux existants [9], [10], [11] .

2. Développement de méthodes de détection automatique de contenus toxiques capables d'exploiter les connaissances modélisées en 1.

Ce deuxième volet des recherches concernera l'augmentation des plongements sémantiques grâce aux annotations en s'appuyant sur des travaux récents qui ont étendu les principes des plongements sémantiques en considérant les représentations des concepts. Ces études utilisent des graphes conceptuels, des ontologies ou taxonomies, et intègrent ainsi les relations conceptuelles définies de manière explicite par ces ressources. Plusieurs travaux ont d'ailleurs étudié comment l'utilisation de relations telles que la synonymie, l'antonymie ou l'hyperonymie permet d'améliorer la qualité des plongements sémantiques, c'est-à-dire la précision de la sémantique capturée par les représentations vectorielles [14].

Ce deuxième volet de recherches abordera également l'utilisation de la génération augmentée de récupération (Retrieval Augmented Generation - RAG) afin de faciliter l'exploration des dimensions sémantiques principalement visées (émotions et engagement du locuteur) dans les corpus. Cette technique optimise les résultats fournis par un grand modèle de langue (Large Language Model – LLM) en faisant appel à une base de connaissances externe aux sources de données ayant été utilisées pour entraîner le modèle avant de générer une réponse. L'utilisation d'une ontologie peut alors améliorer la pertinence des réponses fournies par un LLM [12].

Tandis que les LLMs et les plongements sémantiques sont construits à partir de connaissances génériques, l'objectif de la thèse est donc d'intégrer des connaissances spécifiques, issues d'ontologies et d'annotations linguistiques pour le traitement des contenus toxiques collectés en ligne. La thèse étudiera ensuite l'impact de cette intégration sur les performances des modèles. Les recherches permettront de développer des méthodes d'intelligence artificielle hybrides, au carrefour de la représentation des connaissances et de l'apprentissage automatique [13] ainsi que dans une réflexion renouvelée sur l'apport des connaissances linguistiques aux systèmes de détection de discours de haine [15].

Bibliographie
---

[1] Schafer, Johannes et Kistner, Elina. HS-EMO: Analyzing Emotions in Hate Speech. In : Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023). 2023. p. 165-173.
[2] Frimer, Jeremy A., Brandt, Mark J., Melton, Zachary, et al. Extremists on the left and right

use angry, negative language. Personality and Social Psychology Bulletin, 2019, vol. 45, no 8, p. 1216-1231.
[3] Berger, John M. Promoting disengagement from violent extremism.
[4] Vosoughi, Soroush, Roy, Deb & Aral, Sinan. The spread of true and false news online. Science. 2018 Mar 9;359(6380):1146-1151.
[5] Klinger, Roman. 2023. Bridging emotion role labeling and appraisal-based emotion analysis. arXivpreprint arXiv:2309.02092.
[6] Etienne Aline, Battistelli Delphine, Lecorvé Gwénolé (2024) - "Emotion Identification for French in Written Texts: Considering their Modes of Expression as a Step Towards Text Complexity Analysis", in Actes WASSA - 2024 (14th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis), at ACL 2024 (The 62nd Annual Meeting of the Association for Computational Linguistics), Bangkok, Thailand, August 11–16, 2024).
[7] Battistelli, Delphine, Dragos, Valentina, & Mekki, Jade. 2023b. Annotating social data with speaker/user engagement. illustration on online hate characterization in French. In International Conference on Computing and Communication Networks 2023: ICCCN 2023
[8] Bruneau, Cyril, & Battistelli, Delphine (2024). Guide d'annotation manuelle de la prise en charge énonciative-Version V2 (Février 2024). https://hal.science/hal-04541398/document
[9] Dragos, Valentina, Battistelli, Delphine, Etienne Aline, & Constable,Yolène. (2022, June). Angry or Sad? Emotion Annotation for Extremist Content Characterization. In 13th Language Resources and Evaluation Conference.
[10] Battistelli, Delphine., Bruneau, Cyril., & Dragos, Valentina. (2020). Building a formal model for hate detection in French corpora. Procedia Computer Science, 176, 2358-2365.
[11] Dragos, Valentina, Battistelli, Delphine, & Kelodjoue, Emanuelle (2018, July). Beyond sentiments and opinions: exploring social media with appraisal categories. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 1851-1858). IEEE
[12] Allemang, Dean, & Sequeda, Juan (2024, November). Increasing the Accuracy of LLM Question-Answering Systems with Ontologies. In International Semantic Web Conference (pp. 324-339). Cham: Springer Nature Switzerland.
[13] Antoun, Wissam., Kulumba, Franci., Touchent, Rian., de la Clergerie, Éric., Sagot, Benoît., & Seddah, Djamé. (2024). CamemBERT 2.0: A Smarter French Language Model Aged to Perfection. arXiv preprint arXiv:2411.08868.
[14] Cheng, Siyuan., Zhang, Ningyu., Tian, Bozhong., Chen, Xi., Liu, Qingbin., & Chen, Huajun (2024, March). Editing language model-based knowledge graph embeddings. In Proceedings of the AAAI conference on artificial intelligence (Vol. 38, No. 16, pp. 17835-17843).
[15] Benamara, Farah, Battistelli Delphine, Patti, Viviana (Eds.) (2025). Numéro spécial Discours de Haine : ressources linguistiques, méthodes et applications. Revue TAL (Traitement Automatique des Langues), 65-3. à paraître en mai 2025.

Profil recherché

- Eq. Master en Traitement Automatique des Langues (TAL) ou Ingénierie des Connaissances (IC) ou Ecole d'ingénieurs (maths/info)
- Bonnes compétences en programmation, notamment sous Python
- Des connaissances en annotation linguistique sémantique seront un vrai plus

Modalités de candidature

Encadrantes :

Valentina DRAGOS (Département Traitement de l'information et Systèmes DTIS, ONERA, Palaiseau)
Email : valentina.dragos@onera.fr

Delphine BATTISTELLI (Modyco, UMR 7114 CNRS & Université Paris Nanterre)
Email : delphine.battistelli@parisnanterre.fr

Dossier de candidature à envoyer au plus tard le 15 avril 2025 aux deux encadrantes ci-dessus désignées :
CV, lettre de motivation, relevés de notes du M1 et du M2, le nom d'un.e référent.e à contacter