www.onera.fr

ONERA
THE FRENCH AEROSPACE LAB

PROPOSITION DE STAGE EN COURS D'ETUDES

Référence : **DTIS-2025-03**
(à rappeler dans toute correspondance)

Lieu : Palaiseau

Département/Dir./Serv. : DTIS/MIDL

Tél. : 06 28 45 54 45

Responsable(s) du stage : Valentina DRAGOS

Email. : valentina.dragos@onera.fr

DESCRIPTION DU STAGE

Thématique(s) : Intelligence Artificielle et Décision

Type de stage : ☒ Fin d'études bac+5 ☒ Master 2 ☐ Bac+2 à bac+4 ☐ Autres

Intitulé : Leveraging linguistic annotations to improve automatic detection of online content

Sujet : etection of specific online content remains a persisting issue in social media analysis.This includes extremism, online hate or offensive language, which are topics of interest for regulation of social platforms. Being complex phenomena, automatic methods have to consider a variety of features to capture their nature [1, 2]. Typically, automatic approaches for content detection are based on natural language processing and rely upon both fine-tuning using a specialized dataset and development of learning models [3].

The objective of this work is to broaden the scope of those methods by taking into account knowledge from linguistic annotations, including emotions, appraisals and user engagement [4,7].

Datasets to be used were collected on French social platforms and highlight three types of online content: extremist, hateful and sexist. There are two layers of manual annotation available: a manual annotation of emotions and user engagement. In addition, there also layers of automatic annotations including emotion annotation carried out with two different tools and the annotation of appraisal categories.

The interplay of different linguistic annotations with hate, extremist or sexist content requires further clarification and several experiments will be designed to explore it. To consider both annotations and content type, both features can be learned in a joint model [5]. This work is intended to investigate whether this learning approach benefits from using enriched datasets that contain annotations for content types and linguistic dimensions. To this end, an experimental protocol will be adopted that first focuses on analyzing the correlation of linguistic annotations and content types, i.e. extremist, hateful and sexist. Then, preliminary experiments will be designed to explore methods for learning the phenomena jointly by leveraging linguistic analysis in online content type detection. The experiments will use the transformer-based pre-trained language model BERT [6] and similar architectures to perform classification for various tasks (extremism, hate or sexism detection, etc.), by gradually integrating the layers of linguistic descriptors.

The work has the following milestones:

State of art and problem analysis (1 month)

Analysis of correlations between linguistic annotations and types of content (1 month)

Implementation and fine-tuning of models (1.5 months)

Set up of the experimental protocol, experimentation and analysis of results (1.5 months)

Internship report (1 month)

References:

1. Schäfer, J., & Kistner, E. (2023). HS-EMO: Analyzing Emotions in Hate Speech. In Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023) (pp. 165-173).

2. Yin, W., Agarwal, V., Jiang, A., Zubiaga, A., & Sastry, N. (2023). Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 17, pp. 902-913).

3. Ljubešić, N., Mozetič, I., & Novak, P. K. (2023). Quantifying the impact of context on the quality of manual hate speech annotation. Natural Language Engineering, 29(6), 1481-1494.

4. Dragos, V., Battistelli, D., Sow, F., & Étienne, A. (2024). Exploring the Emotional Dimension of French Online Toxic Content. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 6945-6954).

5. Rabiul Awal, M., Cao, R., Ka-Wei Lee, R., & Mitrovic, S. (2021). AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. arXiv e-prints, arXiv-2103.

6. Wang, Z., Mayhew, S., & Roth, D. (2020). Extending multilingual BERT to low-resource languages. arXiv preprint arXiv:2004.13640.

7. Dragos, V., Battistelli, D., Etienne, A., & Constable, Y. (2022). Angry or Sad? Emotion Annotation for Extremist Content Characterization. In 13th Language Resources and Evaluation Conference .

Est-il possible d'envisager un travail en binôme ? **Non**

Méthodes à mettre en oeuvre :

☐ Recherche théorique	☐ Travail de synthèse
☒ Recherche appliquée	☒ Travail de documentation
☒ Recherche expérimentale	☒ Participation à une réalisation

Possibilité de prolongation en thèse : **Non**

Durée du stage : Minimum : 4 mois Maximum : 6 mois

Période souhaitée : Mars Juilet 2025

PROFIL DU STAGIAIRE

Connaissances et niveau requis :	Ecoles ou établissements souhaités :
Programmation Python, connaissances en ingénierie des connaissances (OWL, RDF) ainsi que traitement du langage naturel serait un avantag	Master 2, Fin d'études BAC +5