3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Postdoctoral research topic

- Title of the proposed topic: Natural language counter-argumentation against online hate speech
- Research axis of the 3IA:
  - Axes 4: AI FOR SMART AND SECURE TERRITORIES
  - Axes 1: CORE ELEMENTS OF AI
- **Supervisor (name, affiliation, email): Serena Villata (Université Côte d'Azur, CNRS, Inria, I3S), email: serena.villata@univ-cotedazur.fr**
- Potential co-supervisor (name, affiliation): Elena Cabrio (Université Côte d'Azur, CNRS, Inria, I3S), email: elena.cabrio@univ-cotedazur.fr
- The laboratory and/or research group: WIMMICS (http://wimmics.inria.fr/) is a research team of Université Côte d'Azur (UCA), Inria, CNRS. The research fields of the team are graph-oriented knowledge representation, reasoning and operationalization to model and support actors, actions and interactions in web-based epistemic communities.

Apply by sending an email directly to the supervisor.
The application will include:

- Letter of recommendation of the supervisors indicated above
- Curriculum vitæ including the list of the scientific publications
- Motivation letter
- Letter of recommendation of the thesis supervisor

- Description of the topic:

Cyber-spaces are important places of exchange, but also places of risk and violence, with consequences for individuals but also for communities in terms of social cohesion. There are four types of prevention: legal, educational, technological and communicative. No single method implemented in isolation gives convincing results and it is important to combine approaches for greater effectiveness. There are already tools that use AI and allow the automatic detection of violent, racist speech, insults, but most of these tools do not integrate cultural and contextual dimensions, and do not go beyond the identification of hate speech, classifying sentences as containing hate speech or not. Using natural language argumentation to constitute a counter-argumentation would be highly beneficial [2]. Argument(ation) mining [3], the new and rapidly growing area of Natural Language Processing (NLP) and

computational models of argument, aims at the automatic recognition of argument structures in large resources of natural language texts.

The **goal** of this post-doc position is to automatically generate counter-arguments against this harmful content. More precisely, the goal is to generate a counter-argument to develop the critical thinking skills of the victims who are the targets of hate messages. The standard approach used on social networks to prevent the spread of hatred is the suspension of users' accounts or the deletion of hateful comments. Our proposal to automatically generate counter-argumentation helps to preserve the right to freedom of expression, counteracting stereotypes with evidence. It can also change the views of harassers, encouraging the exchange of opinions and mutual understanding, and can help to deactivate the hateful content in the conversation.

The **main objectives** of the post-doc program therefore are:

1. Generate natural language counter-arguments, starting incrementally from the generation of claims countering the hate content towards the generation of full argumentative structures composed by evidence and the supported claim. Such counter-arguments will concentrate on the following categories of hate speech: body shaming, ethnicity, sexism and religion.

2. Evaluate the quality of the generated counter-arguments. These arguments will be assessed with respect to the lexical diversity and semantic diversity, but also with respect to the cogency, effectiveness and reasonableness dimensions. In this application scenario, generating "good" arguments is fundamental to assure the effectiveness of the counter-argumentation.

References

[1] Elena Cabrio, Serena Villata. Five Years of Argument Mining: a Data-driven Analysis. Proceedings of 27th International Joint Conference on Artificial Intelligence (IJCAI 2018), pages 5427-5433.

[2] Serra Sinem Tekiroglu, Yi-Ling Chung, Marco Guerini. Generating Counter Narratives against Online Hate Speech: Data and Strategies. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pages 1177-1190.

Keywords:
Natural Language Processing, Hate Speech Detection, Argument Mining, Counter-argumentation, Natural Language Generation

Skills and profile:
• Master degree in Data Science, Computer Science or Computational Linguistics is required.

- Programming skills are required.
- Knowledge of Natural Language Processing and Machine Learning is preferred.
- Fluent English required, both oral and written. French is appreciated but not mandatory.