3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: **Argument-based counter narratives generation to fight online hate speech**
- Research axis of the 3IA:
  Axis 1: Core elements of AI (main axis)
  Axis 4: AI for Smart and Secure Territories (secondary axis)
- **Supervisor (name, affiliation, email): Elena CABRIO (Université Côte d'Azur, Inria, CNRS, I3S), elena.cabrio@univ-cotedazur.fr**
- Co-supervisor (name, affiliation): Serena VILLATA (Université Côte d'Azur, Inria, CNRS, I3S), villata@i3s.unice.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 supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

⇨ **All the requested documents must be gathered and concatenated in a single PDF file named in the following format: LAST NAME of the candidate_Last Name of the supervisor_2023.pdf**

- **Description of the topic:**

Social media have faced mounting pressure from civil rights groups to ramp up their enforcement of anti-hate speech policies. But the increasing availability of online user-generated content and growth of social media platforms present special challenges when it

comes to monitoring and limiting the presence of aggressive and abusive language online[1]. This has led to the deployment of automatic systems – relying on advances in machine learning for Natural Language Processing (under the big umbrella of Artificial Intelligence) - that "understand" expressions in natural language and label the messages as appropriate or not. Computational techniques are necessary to scale up to match a volume of data that would be otherwise impossible to track. Research work on the automatic detection of abusive and offensive language is highlighted by the success of evaluation campaigns such as HatEval, OffensEval, or Automatic Misogyny Identification, to mention a few [1].
The majority of approaches developed so far to automatically detect abusive phenomena are based on supervised machine learning, employing both feature-based classifiers and neural architectures [2]. Specialised lexicons have also been proposed, as a complementary approach to supervised learning with the potential for an increased explainability of the automated decisions of the models.

After detecting it, tackling online hatred using informed textual responses - called *counter narratives* - has been brought under the spotlight recently [3,4]. The automatic generation of counter narratives aims at facilitating the direct intervention in the hate discussion and to prevent hate content from further spreading. While most of the current neural approaches lack grounded and up-to-date evidence such as facts, statistics, or examples, these aspects are of utmost importance and need to be explored to provide convincing and well-grounded arguments. As language models have been successfully employed in the NL generation task [6], we will investigate large-scale unsupervised language models to generate argument-based counter-narratives conditioned on the outputs of an abusive language detection system (able to detect claims in the target abusive message). The generated counter-narrative needs to meet two criteria, namely quality (i.e., variability of the counter-arguments, no repetitiveness) and quantity. In particular, specific strategies will be developed to generate narratives to counter forms of abusive behavior that are linguistically subtle and implicit, that still constitute a real challenge [5].

To sum up, given that we agree that counter narratives are a better tool than content moderation in fighting hate speech, the PhD program proposes to tackle the counter narrative generation task, in particular to undermine the impact of implicit hateful content with informed and non-aggressive responses. Investigating and expanding the scope of problems to tackle both more subtle and more serious forms of abuse aim at promoting healthy online communities.

Keywords:
Natural Language Processing, Machine Learning, Argument Mining, 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.

[1] « Abusive language » includes any expression that uses harsh vocabulary, insults, or more subtle devices such as analogies and stereotypes, to offend, denigrate, or generally cause harm to the recipient of the message.

• Fluent English required, both oral and written. French is appreciated but not mandatory.

References :

[1] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Viviana Patti: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Evaluation 55(2): 477-523 (2021)

[2] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata: A Multilingual Evaluation for Online Hate Speech Detection. ACM Trans. Internet Techn. 20(2): 10:1-10:22 (2020)

[3] Yi-Ling Chung, Serra Sinem Tekiroglu, Marco Guerini: Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. CoRRabs/2106.11783 (2021)

[4] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, Marco Guerini: Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. EMNLP 2022: 8031-8049

[5] Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. EACL 2023.