3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: **Abusive language detection and effective countering**
- 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: MARIANNE (https://team.inria.fr/marianne/). The **MARIANNE** project-team pursues high-impact research in Artificial Intelligence with a focus on data and models for computational argumentation in natural language. MARIANNE is an Inria joint project-team with the I3S (Computer Science) laboratory of Université Côte d'Azur and CNRS. The team is composed of computer scientists, but it holds a strong interdisciplinary connotation in particular with linguistics, philosophy, sociology and law. MARIANNE proposes innovative Natural Language Processing methods, addressing real-world problems, that are both theoretically sound and explainable. The team focuses on topics such as argument mining, argument and counter-argument generation, argument quality assessment, argument-based explainable AI, argument dynamics, argument-based neuro-symbolic models. The main application scenarios investigated in the team are political debates (propaganda), healthcare, law and online social media (hate speech and disinformation).

Apply by sending an email directly to the supervisor.
The application should 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.

- **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].
Research on automatically detecting abusive content in social media relying on Natural Language Processing methods has largely concentrated on overt forms of hate speech, which are easier to identify due to the presence of explicit hateful language [7]. In contrast, messages containing subtle and implicit forms of hate speech - such as circumlocution, metaphors, and sarcasm - pose a significant challenge for automatic detection systems [5,6]. While often explicit hate speech is not argumentative at all, arguments containing implicit and subtle abusive content mostly are, and can be found in political speeches (in particular with the rise of right-wing national populism), or in comments in online forum or social media.
Moreover, when people choose to respond to hateful speech instead of just ignoring it, they often have a variety of motivations and the overarching goal of improving online discourse. To scale, the automatic generation of such informed textual responses - called *counter narratives* - have been brought under the spotlight recently [2,3]. 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 [1].
This PhD program will propose to explore advanced methods to detect implicit offensive content, both as implied negative stereotypes or negatively polarized knowledge (modeled as missing arguments components or fallacious reasoning). Moreover, with the goal of clearing up misunderstandings, reduce tension and setting a positive tone for the discussion, we will improve the effectiveness of current counterargument generation methods across ideologies [4], focusing on fine-grained strategies to attack either specific arguments components, enthymemes or fallacious reasoning. In this context, the dynamic nature of the exchanges and of the debates should be carefully considered, so that to provide effective counterarguments following the flow of the discussion (instead of more general ones, as done by current systems). Moreover, methods to evaluate different impacts of counterspeech by analyzing haters reactions will be investigated, so that to generate effective counter-arguments that will not provoke further negative behavior of the haters [8]. 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:
Artificial Intelligence, Natural Language Processing, Large Language Models, Argument Mining, Abusive language detection, Counter-argument generation

Skills and profile:
- Master degree in Artificial Intelligence, 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.

[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.

References :

[1] Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, Marco Guerini: Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering. EMNLP 2024: 3446-3463

[2] 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

[3] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

[4] Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. Improving argument effectiveness across ideologies using instruction-tuned large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, pages 4604–4622. Association for Computational Linguistics, 2024.

[5] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In Andreas Vlachos and Isabelle Augenstein, editors, Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023.

[6] Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023.

[7] 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)

[8] Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, Lingzi Hong: Echoes of Discord: Forecasting Hater Reactions to Counterspeech. NAACL (Findings) 2025: 4892-4905