3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: **Subtle abusive language detection.**
- 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**
- Potential 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.

- 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

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.

is highlighted by the success of evaluation campaigns such as HatEval, OffensEval, or Automatic Misogyny Identification, to mention a few.
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 [1]. 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.

Current technology has primarily focused on overt forms of abusive language and hate speech, covering only some phenomena along the whole spectrum of toxic and abusive content, while others are ignored because deemed too difficult or rare [2].

While explicit hate speech is more easily identifiable by recognizing a clearly hateful word or phrase, implicit hate speech employs circumlocution, metaphor, or stereotypes to convey hatred of a particular group, in which hatefulness can be captured only by understanding its overall compositional meanings. Although implicitness has an influence on the human perception of hate speech, the phenomenon is invisible to automatic classifiers [3]. This poses a severe problem for automatic abusive language detection, as it opens doors for more intense hate speech hiding behind the phenomenon of implicitness.

The goal of the PhD program is to address the automatic detection of abusive content focusing in particular on the forms of abusive behavior that are linguistically subtle and implicit, that constitute a real challenge for automatic hate speech detection. While the sly, potentially deceiving nature of implicitness might be perceived as less hateful with respect to the same content expressed clearly, such abuse can still be as emotionally harmful as overt abuse.

Moreover, tackling online hatred using informed textual responses - called counter narratives - has been brought under the spotlight recently [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.
Given that we agree that counter narratives are a better tool than content moderation in fighting hate speech, in the second part of the PhD program, the counter narrative generation task will be tackled, 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 aims 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.

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

References :

[1] 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)

[2] David Jurgens, Libby Hemphill, Eshwar Chandrasekharan: A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. ACL (1) 2019: 3658-3666

[3] Darina Benikova, Michael Wojatzki, Torsten Zesch: What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech. GSCL 2017: 171-179

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