# Biases and explainability of Stance Detection Methods

## Research Internship

## Jean Monnet University and University of Sherbrooke

## 4 to 6 months

### Project

This internship takes place as part of a collaboration between LabHC at Jean Monnet University and the University of Sherbrooke and can give rise to an invited stay at the University of Sherbrooke for the trainee.

### Scientific background

The stance detection task [1,2] aims to classify a piece of text expressed by an author as in favor, supporting or being neutral towards a target entity, topic, or claim. This type of analysis is commonly applied to contentious documents/posts discussing controversial topics, such as the Covid-19 vaccine mandate, particularly prevalent in online social media platforms. To enhance the accuracy of stance detection, several techniques have emerged in the supervised setting, leveraging pre-trained language models like BERT [3,4] or Decoder-only Transformers such as GPT-n models.

However, despite the advancements in stance detection, a critical concern arises from the bias of these models, which has been highlighted in other application domains like toxicity detection [5] and sentiment analysis [6,7]. Surprisingly, the issue of bias in stance detection approaches has received little attention, possibly due to the scarcity of sensitive attribute annotations within existing datasets. A way of reducing this would be to provide a transparent decision process: explainability of stance detection algorithms appears crucial.

This research project aims to address these limitations by thoroughly investigating the biases and explainability of stance detection methods.

### Expected work

Creating Stance Detection Datasets with Sensitive Attributes: Since existing stance detection datasets lack annotations for sensitive attributes like SAE/AAE and gender, we will design and build a new dataset that includes these attributes.

Investigating bias in Stance Detection: To address the issue of bias in stance detection, we will conduct a comprehensive study to determine if existing unsupervised and supervised methods are fair across different demographic groups.

Explainable Stance Detection: Building upon the existing language model approaches, we will explore and propose methods to enhance the explainability of stance detection models. Explainability is crucial, especially when the decisions made by the model could impact individuals or communities.

Evaluation of Explainable Stance Detection: To evaluate the proposed explainable stance detection approaches, we will conduct user studies and gather feedback from human annotators.

**Skills required**

The candidate must have solid skills in Machine Learning (mastery of deep learning frameworks such as PyTorch/TensorFlow), as well as advanced skills in Python and generative models.

**Salary**

In accordance with university salary scales

**Additional information**

**Depending on the location of the candidate, the internship can either take place in France or Canada. Mobility during the duration of the internship is also conceivable.**

The Laboratoire Hubert Curien (https://laboratoirehubertcurien.univ-st-etienne.fr) is a joint research unit (UMR 5516) of the Université Jean Monnet de Saint-Etienne, the Centre National de la Recherche Scientifique (CNRS) and the Institut d'Optique Graduate School. The laboratory's research activities are organized into two scientific departments: Optics, Photonics and Surfaces, and Computing, Security and Imaging. The Data Intelligence team, in which the new recruit will work, specializes in Machine Learning.

The University of Sherbrooke is a French-speaking institution known for its human dimension, innovative operating style, and collaboration with professionals. The University of Sherbrooke welcomes more than 31,700 students from 102 countries and territories around the world. The department consists of 24 professors actively involved in the following research areas: artificial intelligence, bioinformatics, health informatics, human-computer interaction, imaging, and digital media sciences, among others. The NLP laboratory focuses on understanding and generating texts in various disciplines related to social issues. Its work includes the identification and mitigation of hateful content, detection of fake news, analysis of positions expressed on social networks, as well as explainability and reduction of biases in NLP models based on neural architectures.

To apply, please send to antoine.gourru@univ-st-etienne.fr, christine.largeron@univ-st-etienne.fr and amine.trabelsi@usherbrooke.ca. Please send a detailed CV, a cover letter and your most recent transcript.

**References**

[1] Schiller, Benjamin, Johannes Daxenberger, and Iryna Gurevych. "Stance detection benchmark: How robust is your stance detection?." KI-Künstliche Intelligenz (2021): 1-13.
[2] Küçük, Dilek, and Fazli Can. "Stance detection: A survey." ACM Computing Surveys (CSUR) 53.1 (2020): 1-37.
[3] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proc. of NAACL. 2019.
[4] Hema Karande and Rahee Walambe et al. "Stance Detection with {BERT} Embeddings for Credibility Analysis of Information on Social Media" ." arXiv preprint arXiv:2105.10272 (2021).
[5] Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).
[6] Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.
[7] Elazar, Y., & Goldberg, Y. (2018). Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 11-21).