Offer #2025-08767

Post-Doctoral Research Visit F/M Postdoctoral researcher in Responsible AI for Journalism

Level of qualifications required : PhD or equivalent

Fonction : Post-Doctoral Research Visit

About the research centre or Inria department

The Inria Saclay-Île-de-France Research Centre was established in 2008. It has developed as part of the Saclay site in partnership with Paris-Saclay University and with the Institut Polytechnique de Paris since 2021.

The centre has 39 project teams , 27 of which operate jointly with Paris-Saclay University and the Institut Polytechnique de Paris. Its activities occupy over 600 scientists and research and innovation support staff, including 54 different nationalities.

Context

Every year Inria International Relations Department has a few postdoctoral positions in order to support Inria
international collaborations.
The postdoctoral contract will have a duration of 18 months. The start date is between July 1st and September 1st, but not later than September 1st.

Team:

A potential postdoctoral researcher would integrate the Inria CEDAR team while also visiting the Human-Centered Data Analytics team at CWI in Amsterdam. This project is a joined collaboration with the following PIs:

Oana Balalau is an Inria researcher in the team CEDAR, at the Inria center of Institut Polytechnique de Paris. Her research interests are in natural language processing, in particular in argumentation mining, information extraction and data2text. She is collaborating with journalists from several news agencies: Radio France, Le Monde and AEF Info.

Davide Ceolin is a senior CWI researcher in the Human-Centered Data Analytics group. His research focuses on transparently predicting multiple aspects of information quality. He is a member of the AI, Media, and Democracy lab, a multidisciplinary lab that studies in depth the effects and implications of AI for Media and Democracy. The lab brings together Computer Science, Legal, and Communication scholars, as well as several civil society and industrial partners.

The interested candidates can contact Oana Balalau if they have additional questions (oana.balalau@inria.fr).

Assignment

Candidates for postdoctoral positions are recruited after the end of their Ph.D. or after a first post-doctoral
period: for the candidates who obtained their PhD in the Northern hemisphere, the date of the Ph.D. defense
shall be later than September 1, 2022; in the Southern hemisphere, later than April 1, 2022.
In order to encourage mobility, the postdoctoral position must take place in a scientific environment that is truly
different from the one of the Ph.D. (and, if applicable, from the position held since the Ph.D.); particular attention
is thus paid to French or international candidates who obtained their doctorate abroad.

Context: From recommender systems to large language models, **AI tools** have shown **different forms of limitations and bias [BHA+21, MMS+21, NFG+20]**. **Bias in AI tools** may stem from multiple factors, including bias in the input data the AI tools are trained on, the algorithm and the individuals responsible for designing the AI tools, and bias in the evaluation and interpretation of AI tool outputs [NFG+20]. **Limitations** are due to technical difficulties in achieving specific tasks [SB22]. Media outlets use different algorithmic aids in their workflow: entities and relations extractions, event extraction, sentiment analysis, automatic summarization, newsworthy story detection, semi-automatic production of news using text generation models, and AI-guided search, among others [TJM+22, UBM23]. Given the importance of the media sector for our democracies, shortcomings in the tools they use could have severe consequences.

Main activities

Research question:

What are the potential sources of bias in natural language processing (NLP) driven applications targeted for journalism and how can we highlight them and mitigate their effect?

To answer this question, we will investigate two use cases.

Bias and limitations in classification tasks. We have developed a fact-checking platform where journalists can monitor politicians' statements on social media [BEG+22]. Statements that are more likely to be checkworthy are highlighted, and for this, we used a machine learning algorithm. **Checkworthy claims** are defined as factual sentences that the general public will be interested in knowing whether they are true [HAL+17]. We note that this definition is based on what an annotator considers as being of general interest. In addition, the checkworthy training dataset contains political statements. Hence, annotators might have inadvertently introduced political bias in their annotations, for example, by labeling sentences more often as checkworthy if they are expressed by someone of a different political affiliation than their own. A second model used in our pipeline is detecting **propaganda**, where propaganda is defined as a set of communication techniques that are designed to influence a reader, not to inform them. Of particular interest are fallacious arguments, which are incorrect arguments that fact-checkers should debunk. While propaganda definitions are more precise depending on the exact type of technique (e.g., loaded language, ad hominem), annotated datasets often have low inter-annotator agreement [DSB+19]. In addition, the datasets also contain only political statements - again, an annotator could be more inclined to label the speech of someone of a different political view as propaganda. We would like to investigate if such datasets and models are biased, and if this is the case, investigate how it could be possible to highlight the bias. One interesting idea is to incorporate disagreement into a classification task by providing a textual explanation of why a certain paragraph could have two or more different labels (also known in ML as multi-label classification) according to two or more different human opinions. As mentioned, the disagreement could come from the definition of the task but also from the beliefs of the annotators. This entails rethinking the annotation process, training and evaluation of an NLP model, and the way a model is used for a real application. We note that the problem of variability and bias in human annotation is getting more attention in the NLP community [P22, UFH+21].

Bias and limitations in generative tasks. Nowadays, generative language models are used for a variety of tasks, in particular for essays or **argumentative texts**. We have discussed this with journalists, who have confirmed they are using such tools to speed up their work. We would like to focus on argumentative texts, particularly on controversial topics in our society. To investigate the potential bias of

argumentative models when asked to provide information on such topics, we would like to compare automatically generated argumentative texts with crowdsourced argumentative texts, such as text hosted on debate platforms. This project can be extended to analyzing how **controversial topics** are debated in the public sphere, for example, by focusing on debates in current electoral races. As a technical challenge for this task, the first one is identifying similar arguments - when an argument is composed of a claim and the evidence supporting the claim. The same claim can be supported by different evidence, and highlighting such differences is also important, as a preference over a certain type of evidence could highlight greater trends. For example, the claim “Abortion should be legal.” can be supported by “A woman should always have the choice over her body.” or the sentence “God has given us free will, and we should respect the free will of others.”. A second technical challenge is in measuring how **persuasive** an argumentative text is, for example, by measuring how complete is the evidence brought forward [HG16].

References:

[BEG+22] Balalau, O., Ebel, S., Galizzi, T., Manolescu, I., Massonnat, Q., Deiana, A., Gautreau, E., Krempf, A., Pontillon, T., Roux, G. and Yakin, J., 2022, October. Fact-checking Multidimensional Statistic Claims in French. In *TTO 2022-Truth and Trust Online*.

[BHA+21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[DSB+19] Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Petrov, R. and Nakov, P., 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5636-5646). Association for Computational Linguistics.

[HAL+17] Hassan, N., Arslan, F., Li, C. and Tremayne, M., 2017, August. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1803-1812).

[HG16] Habernal, I. and Gurevych, I., 2016, November. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1214-1223).

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*,54(6):1–35, 2021.

[NFG+20] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356,2020.

[P22] Plank, B., 2022, December. The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10671-10682).

[SB22] Chirag Shah and Emily M Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 221–232, 2022.

[TJM+22] Christoph Trattner, Dietmar Jannach, Enrico Motta, Irene Costera Meijer, Nicholas Diakopoulos, Mehdi Elahi, Andreas L Opdahl, Bjørnar Tessem, Nj?al Borch, Morten Fjeld, et al. Responsible media technology and ai: challenges and research directions. *AI and Ethics*, *2*(4):585–594, 2022.

[UBM23] Prajna Upadhyay, Oana Balalau, and Ioana Manolescu. Open information extraction with entity focused constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1255–1266, 2023.

[UFH+21] Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B. and Poesio, M., 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, pp.1385-1470.

Skills

Technical skills and level required : strong knowledge of NLP and good programming skills

Languages : English

Benefits package

- Subsidized meals
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)

- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training

Remuneration

According to profile

General Information

- **Theme/Domain :** Data and Knowledge Representation and Processing Statistics (Big data) (BAP E)
- **Town/city :** Palaiseau
- **Inria Center :** Centre Inria de Saclay
- **Starting date :** 2025-07-01
- **Duration of contract :** 1 year, 6 months
- **Deadline to apply :** 2025-06-30

Contacts

- **Inria Team :** CEDAR
- **Recruiter :**
  Balalau Oana-denisa / oana.balalau@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

The candidate should submit:

- Detailed CV with a description of the PhD and a complete list of publications with the two most significant ones highlighted
- Motivation letter
- 2 letters of recommendations
- Passport copy

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.