Francielle Vargas

PhD in Computer Science - Artificial Intelligence

franciellealvargas@gmail.com

https://franciellevargas.github.io/

Summary

Experienced researcher in Artificiall Intellgience - Natural Language Processing with a focus on Language Model Interpretability. I have made significant contributions to academia, with a strong track record of publications in top-tier NLP venues. My work bridges research and practice, leading and collaborating on international, interdisciplinary projects that deliver real-world impact and drive positive societal change mainly for underrepresented communities.

Education

PhD, Computer Science and Computational Mathematics
University of São Paulo, Brazil
2019 - 2024
Thesis: Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection.
Artificial Intelligence, Natural Language Processing

MSc, Computer Science and Computational Mathematics
University of São Paulo, Brazil
2015 - 2017
Dissertation: Semantic Clustering of Aspects for Opinion Mining.
Artificial Intelligence, Natural Language Processing

BA, Applied Linguistics
Federal University of Minas Gerais, Brazil
2010 - 2014

BS, Information Systems
Pontifical Catholic University of Minas Gerais, Brazil
2006 - 2009

Publications

The complete list of publications: https://franciellevargas.github.io/

Conference Participation / Presentation

- 2025: EMNLP, CSBC
- 2024: EMNLP, NAACL
- 2023: ACL, RANLP
- 2022: LREC
- 2021: RANLP, ICWSM
- 2020: ACL, EMNLP, LREC, BRACIS

Experience

04-2025 to 03-2025: **Postdoctoral Researcher, São Paulo State University, Brazil & Idiap Research Institute, Switzerland.**
The RATIONAL research project is an international collaboration between the Idiap Research Institute in Switzerland and São Paulo State University in Brazil, with research teams based in both countries. My contribution focuses on the development of robust, evidence-based retrieval methods and interpretable approaches for Natural Language Inference over Transformer-based models, supported by LLMs.

01-2024 to 12-2024: **Google LARA PhD Fellowship, Brazil**
Awarded the Google PhD Fellowship to support my doctoral research in the field of Trustworthy AI, focusing on automated fact-checking, and Language Model Interpretability. This prestigious fellowship provided the opportunity to engage with leading researchers in the AI and NLP field and attend prestigious venues worldwide.

04-2024 to 04-2024: **Visiting PHD Student, University of Southern California, USA**
Collaborated with researchers at USC. I focused on improving model interpretability and fairness to the detection and mitigation of hate speech. As a result, we submitted a paper at EMNLP 2025 that proposes a method based on multi-hop hate speech explanation for moral reasoning evaluation of Large Language Models (LLMs).

11-2023 to 11-2023: **Invited Researcher, Leibniz Institute for the Social Sciences, Germany**
I was honored to be invited as a guest researcher to speak at the Conference on Harmful Online Communication (CHOC2023). During the event, I shared insights from my research on disinformation, hate speech detection, and the ethical implications of AI in marginalized communities, contributing to important discussions on mitigating harm in digital communication.

08-2021 to 12-2021: **Teaching Assistant - Neural Networks and Deep Learning, University of São Paulo, Brazil**
Assisted in teaching a graduate-level course on Neural Networks and Deep Learning. Responsibilities included preparing lecture materials, grading assignments, conducting laboratory sessions, and mentoring students on the application of deep learning techniques.

08-2020 to 12-2020: **Teaching Assistant - Computing Theory and Compilers, University of São Paulo, Brazil**
Assisted in teaching an undergraduate-level course on Computing Theory and Compilers. Responsibilities included preparing instructional materials, grading exams and assignments, supporting laboratory and problem-solving sessions, and assisting students in understanding formal models of computation and compiler design concepts.

07-2019 to 12-2023: **Doctoral Researcher, University of São Paulo, Brazil**
Obtained a Ph.D. in Artificial Intelligence (Natural Language Processing), with a focus on AI for social impact. My dissertation focused on language model interpretability and fairness, contributing a wide range of benchmark datasets as well as post-hoc and self-explaining computational methods for the detection of hate speech and disinformation in underrepresented communities in the Global South. I collaborated with several international institutions and published research at top-tier NLP conferences. My thesis has been recognized with both national and international awards.

Awards, Honors and Grants

1. **International Trevisan Prize for Students "AI for Good", Bocconi University, Italy.**
   In 2026, I received the inaugural Trevisan Prize for Students "AI for Good" for my Ph.D. dissertation, "Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection," which advances fairness, accountability, and explainability in NLP, particularly for automated fact-checking and hate speech detection in underrepresented languages. The biennial prize honors the legacy of Prof. Dr. Luca Trevisan by recognizing rigorous computer science research and outstanding contributions to AI with the potential for significant, positive, and long-lasting societal impact.
2. **Maria Carolina Monard Award for the Best Thesis in Artificial Intelligence, University of São Paulo, Brazil.**
   In 2025, my Ph.D. thesis was recognized as the best in Artificial Intelligence. The Maria Carolina Monard Award annually recognizes the best PhD thesis in Computer Science in the field of Artificial Intelligence in Brazil, highlighting originality, scientific, technological, cultural, and social relevance, and innovation potential.
3. **Nominated by the Brazilian Computer Society for the Thesis Award in the broader Computer Science area**
   In 2025, my PhD thesis was selected as a finalist in the Brazilian Computer Society's Thesis and Dissertation Competition (CTD), one of the most prestigious and competitive awards for graduate research in computer science in Brazil. Given the country's size and the large number of graduate programs, this recognition is highly selective and reflects the scientific excellence, originality, and potential impact of the work. I was honored to be among the top 11 Ph.D. Theses nationwide.
4. **Nominated by the Brazilian Computer Society for the Thesis Award in Multimedia, Hypermedia, and the Web**
   In 2025, my Ph.D. thesis was selected among the top 6 best theses in the country in the fields of Multimedia, Hypermedia, and the Web, being recognized as a finalist for the Thesis and Dissertation Award (CTD) at the Brazilian Symposium on Multimedia and the Web (WebMedia).
5. **Google Latin America Research Award (LARA)**
   In 2024, my PhD research project was awarded the Google Latin America Research Award (LARA) as part of a larger research initiative on combating misinformation in Latin America, led by my co-advisor, Professor Dr. Fabrício Benevenuto. The LARA Google PhD Fellowship is designed to support innovative research in various fields of computer science, including artificial intelligence, machine learning, and natural language processing. The awards aim to support researchers and faculty members based in Latin America who are conducting cutting-edge research with the potential for significant impact in their respective fields.
6. **Awarded the ACL Grants (EMNLP & NAACL)**
   In 2024, I received a Diversity & Inclusion (D&I) grant from the Association for Computational Linguistics (ACL), which provided conference registration support to attend EMNLP and NAACL 2024. This award is granted to Ph.D. researchers in recognition of their outstanding contributions and achievements in Natural Language Processing and Computational Linguistics, as well as to applicants from underrepresented groups presenting a paper at the main conference.
7. **Outstanding Academic Achievement with Honorable Mention, Federal University of Minas Gerais (UFMG)**

In 2012 and 2013, for two consecutive years, my undergraduate research projects received an award for academic relevance and honorable mention during my studies at the UFMG. This award recognizes the best research projects across all undergraduate programs at the university for that year.

Invited Speaker

2024: **Invited Talk at the Computational Social Science - Language and Morality Lab, University of Southern California (USC), Los Angeles, CA**.
Talk: Fighting Misinformation and Radicalism: Socially Responsible and Explainable Fact-Checking and Hate Speech Detection.
During my visit to USC, I was invited by Professor Morteza Dehghani to speak at the Computational Social Science - Language and Morality Lab. I presented the methods and benchmarks that we developed in Brazil to improve explainability and fairness in fact-checking and hate speech detection.

2024: **Keynote Speaker at the Conference cum Conclave on Emerging trends in Journalistic and Media Practices, DG Vaishnav College (DDGDV), Chennai, India**.
Talk: Predicting Sentence-Level News Source Reliability for Fact-Checking.
I was invited as a Keynote Speaker at this conference, organized by DG Vaishnav College (DDGDV), India. I presented insights and findings from my paper published in RANLP 2023 "Predicting Sentence-Level Factuality of News and Bias of Media Outlets," discussing methods for assessing news reliability and media bias.

2023: **Keynote Speaker at the Conference on Harmful Online Communication, Leibniz Institute for the Social Sciences (GESIS), Cologne, Germany**.
Talk: Countering Harmful Online Communication in Brazil: Predicting Fine-Grained Factuality of News and Offensive Context of Social Media Comment.
I was invited as a Keynote Speaker by the conference organizers, to speak alongside prestigious researchers, including Isabelle Augenstein (University of Copenhagen), Leon Derczynski (University of Washington), and Libby Hemphill (University of Michigan), among others. The conference discussed methods to address harmful speech worldwide. My talk focused on advancing explainability and fairness in fact-checking and hate speech detection.

Organizing Committee

- **Co-Organizer, DeepXplain @ IJCNN 2025**: As the lead co-organizer, I collaborated with Professor Dra. Roseli Romero, an associate professor in the Department of Computer Science at the University of São Paulo (USP), Brazil, and Dr. Jackson Trager, a social psychologist specializing in ethics at the University of Southern California, USA, to plan and execute the Special Session on Explainable Deep Neural Networks for Responsible AI at the International Joint Conference on Neural Networks (IJCNN 2025), which will take place in Rome, Italy. This workshop focuses on explainable deep neural networks, aiming to advance trustworthy AI practices.
- **Co-Organizer, WOAH @ ACL 2025**: I was invited to be part of the organizing committee for the 9th Workshop on Online Abuse and Harms (WOAH), co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), which will take place in Vienna, Austria. I helped organize this workshop, which focuses on tackling harmful online communication and its implications, bringing together researchers working on online abuse and its detection.

- **Co-Organizer, ICWSM 2023, 2022, 2021**: I served as co-chair of the Datasets Track, and as Diversity, Inclusion, and Accessibility Chair at the International AAAI Conference on Web and Social Media (ICWSM) in 2021, 2022, and 2023. I managed submissions and ensured that high-quality datasets were presented for machine learning and web mining research. In addition, I promoted a more inclusive environment by improving accessibility for attendees with disabilities, including content availability and communication support.

Program Committee

1. **Journal Reviewer**: I am a reviewer for some of the most prestigious international journals in Natural Language Processing (NLP), including:
   - Natural Language Engineering, Cambridge University Press.
   - Language Resources and Evaluation, ELRA.
   - Online Social Networks and Media.
2. **Conference Reviewer**: I am a reviewer for the main international conferences in Natural Language Processing (NLP) and Machine Learning (ML), including:
   - Empirical Methods in Natural Language Processing (EMNLP)
   - Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)
   - Annual Meeting of the Association for Computational Linguistics (ACL)
   - International Conference on Language Resources and Evaluation (LREC)
   - International Conference on Computational Linguistics (COLING)
   - International AAAI Conference on Web and Social Media (ICWSM)
   - Conference on Information and Knowledge Management (CIKM)

Mentoring

- **Master's Degree Student in Computer Science, Federal University of Minas Gerais, Brazil (2024)**: I co-advised Isadora Salles in developing the first benchmark dataset for explainable hate speech detection in Brazilian Portuguese, named HateBRXplain. This work was published at COLING 2025, a top-tier NLP conference. The dataset is available in https://github.com/isadorasalles/HateBRXplain.
- **Undergraduate Student in Computer Science, University of São Paulo, Brazil (2020)**: I co-advised Lucas Sobral Fontes Cardoso on his final project, which proposed a new framework for opinion extraction and clustering from web consumer reviews. As a result, we developed a new opinion mining system for Portuguese, available at: http://www.nilc.icmc.usp.br/opcluster/.

Benchmark Datasets

HateBRXplain and HateBRMoralXplain: Benchmark Datasets with Hate Speech and Moral Human-Annotated Rationales for Explainable Hate Speech Detection.
MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of Large Language Models
HateBR: A Benchmark Dataset for Explainable Hate Speech Detection in Brazilian Portuguese.
FactNews: A Benchmark Dataset for Sentence-Level Factuality and Media Bias Prediction.

HausaHate: A Benchmark Expert Dataset for Hausa Hate Speech Detection.
MOL: A Context-Aware Multilingual Offensive Lexicon.

Computational Methods

SELFAR: Sentence-Level Factual Reasoning for Explainable Fact-Checking.
SRA: Supervised Rational Attention for Self-Explaining Hate Speech Detection.
SMRA: Supervised Moral Rational Attention for Self-Explaining Hate Speech Detection.
B+M: Contextual BoW with Feature Saliency for Explainable HS Detection.
SSA: Counterfactual Explanations to Assess Social Bias in Hate Speech Classifiers.

Computer Programs

FACTual: A Fact-Checking and News Source Reliability Estimation System
NoHateBrazil: A Brazilian Portuguese Text Offensiveness Analysis System
OPCluster-PT: A System for Extraction and Clustering of Opinions.

Relevant Short Courses (Attended)

2024: 2nd Mexican NLP Summer School, NAACL 2024, Mexico.
2023: 2nd Summer School on Deep Learning in NLP RANLP 2023, Bulgaria.
2023: IEEE Spoken Language Technology Workshop Hackathon, Qatar.
2022: 2nd Advanced NLP School, Université Grenoble Alpes, France.
2021: 4th Advanced School in Big Data Analysis, ICMC-USP, Brazil.
2020: 10th Lisbon Machine Learning School, INESC, Portugal.
2020: Hackathon Antisemitism on Social Media, Indiana University, USA.
2019: Introduction on the Stars - Astrophysics, University of São Paulo, Brazil.

References

Dr. Ameeta Agrawal. Assistant Professor of the Department of Computer Science, Portland State University, USA.
Email: ameeta@pdx.edu.
I have established a successful and productive international collaboration with Dr. Agrawal, which has resulted in several papers published and submitted to EMNLP 2024 and 2025, ACL 2026, and TACL. Our research addresses key topics such as explainability and interpretability, bias mitigation, and factuality prediction, and moral reasoning evaluation of LLMs making a significant contribution to the advancement of the field.

Dr. Roseli Romero. Associate Professor of Computer Science at the University of São Paulo, Brazil.
Email: rafrance@icmc.usp.br.
I had the privilege of working as a teaching assistant under the supervision of Dr. Roseli Romero for the Neural Networks and Deep Learning course in 2021. Additionally, Dr. Romero and I, in collaboration with Dr. Jackson Trager from the University of Southern California, organized the special session Deep Neural Networks for Responsible AI (DeepXplain), co-located with IJCNN 2025 in Italy. The workshop aims to foster important discussions and advancements in interpretability, explainability, and the trustworthy use of deep neural networks.

Dr. Debora Nozza. Assistant Professor in the Computer Science Department at Bocconi University, Italy.
Email: debora.nozza@unibocconi.it
I first met Dr. Débora through our shared involvement in the ACL community, particularly in initiatives focused on hate speech and Responsible and Socially-Aware Natural Language Processing. Our collaboration deepened when we worked together on the 9th Workshop on Online Abuse and Harms (WOAH), co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). I joined the organizing

committee she co-chaired after being selected through a highly competitive selection process. Working together on the program, reviewing processes, and shaping the 9th WOAH workshop's thematic focus allowed us to align our shared commitment to advancing research on hate speech and ethical, inclusive AI. Dr. Debora has also been exceptionally generous and highly supportive of my career development, including endorsing my application for the ACL Dissertation Award in 2026.

Dr. Fabricio Benevenuto. Associate Professor of Computer Science at Federal University of Minas Gerais, Brazil.
Email: fabricio@dcc.ufmg.br.
I have had the privilege of working closely with Dr. Fabricio Benevenuto, who was my co-advisor during my PhD. Dr. Benevenuto is an internationally recognized expert in the fields of disinformation and hate speech, consistently ranked among the most influential researchers in the world.

Dr. Morteza Dehghani. Professor of Psychology and Computer Science, Director, Center for Computational Language Sciences, University of Southern California, USA.
Email: mdehghan@usc.edu.
I had the incredible opportunity to visit the MOLA – Morality and Language Lab at the University of Southern California (USC), coordinated by Dr. Morteza Dehghani. My visit focused on interdisciplinary research at the intersection of computational linguistics, psychology, and explainable and responsible AI. Dr. Dehghani's expertise in cognitive modeling, computational social science, and natural language processing greatly enriched my understanding of how social-psychological theories can inform computational models, particularly in the context of Large Language Models (LLMs). As a result of this visit, Dr. Dehghani's students and I submitted a paper to EMNLP 2025, in which we introduce the first multilingual benchmark dataset for evaluating the moral reasoning capabilities of LLMs.

Dr. Eduard Hovy. Executive Director of Melbourne University, Melbourne, Australia, and Associate Professor in the Language Technologies Institute at Carnegie Mellon University, USA.
Emails: ehovy@andrew.cmu.edu and hovy@cmu.edu.
During my Ph.D., I had the privilege of receiving guidance from Dr. Eduardo Hovy, who generously shared valuable insights on topics related to my research. We had the opportunity to meet in person at RANLP 2023 in Bulgaria, where we discussed key aspects relevant to the advancement of my work. I am currently collaborating with his Ph.D. student, Matteo Guida from the University of Melbourne. Our collaboration has resulted in a paper published at EMNLP 2025, a second work currently under submission to ACL 2026, and ongoing joint work toward a journal submission to TACL, focused on monolingual and multilingual expert-annotated benchmarks and self-explaining methods for hate speech and moral reasoning in large language models.