Francielle Vargas

PhD in Computer Science - Artificial Intelligence

franciellealvargas@gmail.com

https://franciellevargas.github.io/

Summary

Experienced researcher in Machine Learning and Natural Language Processing with a focus on Explainability, Bias Mitigation, and Fairness. I have made significant contributions to academia, with a strong track record of publications in top-tier NLP venues. Proven leadership and collaboration on international, interdisciplinary research projects, academic service, and mentoring.

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, Machine Learning

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, Machine Learning

BA, 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

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

Experience

01-2024 to 12-2024: **Google PhD Fellowship, Federal University of Minas Gerais, Brazil**
Awarded the Google PhD Fellowship to support my doctoral research in the field of Trustworthy AI, focusing on disinformation detection, and the explainability of AI systems. This prestigious fellowship provided the opportunity to engage with leading researchers in the AI field and attend conferences and workshops worldwide.

04-2024 to 04-2024: **Visiting Researcher, University of Southern California, USA**
Collaborated with researchers at USC 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 hate speech multi-hop 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, 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 lab sessions, and mentoring students on the application of deep learning techniques.

08-2019 to 12-2023: **Doctoral Researcher, University of São Paulo, Brazil**
Obtained a PhD in Natural Language Processing with a focus on AI for social impact. My dissertation work centered on developing explainable AI models for hate speech and disinformation detection. I collaborated with various international institutions and published research in NLP top-tier conferences.

12-2021 to 04-2022: **Data Scientist, Cisco Webex, USA**
Worked as a data scientist as a contractor at Cisco in the AI and machine learning team. I collaborated in the internationalization project of Webex developing predictive models, enhancing product features with machine learning, and analyzing large datasets to optimize user experience and service quality.

08-2015 to 11-2017: **MSc Researcher, University of São Paulo, Brazil**
Research projects related to natural language processing and machine learning. Focused on improving text classification techniques to real-world problems, such as sentiment analysis and opinion mining.

09-2014 to 08-2015: **System Analyst, Unisys, Brazil**
Developed and maintained software solutions for enterprise clients. Worked on systems analysis, database design, and the implementation of business applications, contributing to the optimization of processes and the delivery of IT solutions.

Awards & Honors

1. **Brazilian Computer Society Thesis Award Finalist (CTD)**
   In May 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. My thesis focused on responsible AI and natural language processing mainly in hate speech detection and explainability. I was honored among the top PhD theses nationwide.

2. **Google Latin America Research Awards (LARA)**
   In January 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.

3. **Diversity and Inclusion Award (ACL)**
   In October 2024, I received the D&I award from the Association for Computational Linguistics (ACL), which provided travel support and EMNLP 2024 conference registration grants to PhD researchers recognized for their exceptional contributions and accomplishments in the field of Natural Language Processing.

4. **Diversity and Inclusion Award (ACL)**
   In May, 2024, I received the D&I award from the Association for Computational Linguistics (ACL), which provided travel support and NAACL 2024 conference registration grants to PhD researchers recognized for their exceptional contributions and accomplishments in the field of Natural Language Processing.

5. **Outstanding Academic Achievement with Honorable Mention, Federal University of Minas Gerais (UFMG)**
   For two consecutive years, in 2012 and 2013, 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 Speaker 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-Organizing 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-Organizing 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-Organizing ICWSM 2023**: I was a co-chair of the datasets track at the 17th International AAAI Conference on Web and Social Media (ICWSM) held in Limassol, Cyprus. I managed submissions and ensured that high-quality, datasets were presented for machine learning, and web mining research.
- **Co-Organizing ICWSM 2022**: As Accessibility Chair, I worked on making the 16th International AAAI Conference on Web and Social Media (ICWSM), held in Atlanta, Georgia, more inclusive by ensuring accessibility for attendees with disabilities, including content availability and communication facilitation.
- **Co-Organizing ICWSM 2021**: As Diversity, Equity, and Inclusion Chair, I led initiatives to promote diversity, equity, and inclusion at the 15th International AAAI Conference on Web and Social Media (ICWSM), ensuring a more inclusive environment for participants from diverse backgrounds. We introduced the first student travel grants, and registration grants for researchers from marginalized groups.

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.
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: A Benchmark Dataset with 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.

Novel Computational Methods

SELFAR: The First Explainable Fact Checking Benchmark for Portuguese.
B+M: Contextualized Bag-of-Words with Feature Saliency for Explainable Hate Speech Detection.
SSA: A Counterfactual Explanation Approach to Assess Social Bias in Hate Speech Classifiers.

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

Dra. Ameeta Agrawal. Assistant Professor of the Department of Computer Science, Portland State University, USA.
Email: ameeta@pdx.edu. (+1) 503 997-6449
I have established a successful and productive international collaboration with Dra. Agrawal, which culminated in the publication of two papers at EMNLP 2025 and 2024. 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. Fabricio Benevenuto. Associate Professor of Computer Science at Federal University of Minas Gerais, Brazil.
Email: fabricio@dcc.ufmg.br. (+55) 31 99319-1584.
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. Eduard Hovy. Executive Director of Melbourne University, Melbourne, Australia, and Associate Professor in 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 via email on topics related to my research. In addition, 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. Currently, I am collaborating with his Ph.D. student, Matteo Guida from University of Melbourne, to propose the first multilingual benchmark dataset annotated by experts with hate speech multi-hop explanation for evaluating the moral reasoning of LLMs.

Dra. Flor Plaza-del-Arco. Assistant Professor Institute of Advanced Computer Science, Leiden University, Netherlands.
Email: f.m.plaza.del.arco@liacs.leidenuniv.nl
I have had the pleasure of working with Dr. Flor Plaza-del-Arco to organize the 9th Workshop on Online Abuse and Harms (WOAH), co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Together with other outstanding researchers, we worked to bring attention to the critical issues of hate speech and misinformation, aiming to foster a space for sharing insights and developing solutions. Additionally, Dra. Flor Plaza-del-Arco is a co-author of the paper "MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of LLMs through Hate Speech Multi-hop Explanations," submitted to EMNLP 2025, in which we propose the first multilingual benchmark for evaluating the moral reasoning of large language models.

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

Dra. Roseli Romero. Associate Professor of Computer Science at the University of São Paulo, Brazil.
Email: rafrance@icmc.usp.br. (+55) 16 99218-0458.
I had the privilege of working as a teaching assistant under the supervision of Dra. 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, 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.