

Francielle Vargas
PhD in Computer Science - Artificial Intelligence
franciellealvargas@gmail.com
<https://franciellealvargas.github.io/>

Summary Experienced researcher in Machine Learning and Natural Language Processing with a focus on Explainability, Bias Mitigation, and Fairness. Strong background in developing socially responsible AI solutions, including explainable fact-checking and hate speech detection. Proven leadership and collaboration in international 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

Short Courses

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.
2016: **The R language with an emphasis on probability**, UFMG, Brazil.
2015: **4th Summer School in Computer Science**, UFMG, Brazil.
2014: **Software Engineering (Specialization)**, Pontifical Catholic University of Minas Gerais, Brazil.

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 on projects related to computational social science and AI ethics. My research focused on applying AI models to the detection and mitigation of disinformation in social media platforms, as well as improving model interpretability and fairness.

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. 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 detection and disinformation analysis. I collaborated with various international institutions and published research in top-tier conferences.

12-2021 to 04-2022: Data Scientist, Cisco Webex, USA

Worked as a data scientist in the AI and machine learning team at Cisco Webex. My role involved developing predictive models for customer engagement, 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

Engaged in research projects related to natural language processing and machine learning. Focused on improving text classification algorithms and applying these techniques to real-world problems, such as sentiment analysis and social media data 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. **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.

2. **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.

3. **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.

4. **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 Federal University of Minas Gerais (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 we developed in Brazil to enhance explainability and fairness in automated fact-checking and hate speech detection.

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 Katrin Weller, Pascal Siegers, Indira Sen, and Christina Dahn, 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 trending 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 WOAHA @ 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, European Language Resources Association.
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)
 - The 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)
-

Conference Participation

- ACL 2023, ACL 2020.
 - EMNLP 2024, EMNLP 2020.
 - LREC 2022, LREC 2020.
 - RANLP 2023, RANLP 2021.
 - NAACL 2024.
 - ICWSM 2021.
-

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.
 - **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

HateBR: A Benchmark Dataset for Explainable Hate Speech Detection in Brazilian Portuguese.
HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection.
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.

Benchmark Models

SELFAR: The First Explainable Fact Checking Benchmark for Portuguese.

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 a paper at EMNLP 2024 FEVER. Our research addresses key topics such as explainability and interpretability, bias mitigation, and factuality prediction, making a significant contribution to the advancement of the field.

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, to propose a new benchmark dataset for explainable hate speech detection.

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 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 workshop 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.

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. Morteza Dehghani. Professor of Psychology and Computer Science, Director, Center for Computational Language Sciences, University of Southern California, USA.

Email: mdehghan@usc.edu

During my visit to the University of Southern California (USC) in 2024, I had the incredible opportunity to collaborate with Dr. Morteza Dehghani and his students. My visit focused on interdisciplinary research at the intersection of computational linguistics, psychology, and responsible AI. Dr. Dehghani's expertise in cognitive modeling and natural language processing greatly enriched my understanding of how psychological theories can inform computational models, particularly in the context of Large Language Models (LLMs).

Dra. Debora Nozza. Assistant Professor in Computing Sciences at Bocconi University, Italy.

Email: debora.nozza@unibocconi.it

I have had the pleasure of working with Dr. Debora Nozza to organize the 9th Workshop on Online Abuse and Harms (WOAH). We worked alongside other respected researchers to bring attention to the critical issue of hate speech, aiming to create a space for researchers to share insights and develop solutions. Furthermore, Dr. Nozza and I met in person at EMNLP 2024 in Miami, USA, where we further discussed and explored new opportunities for future collaborations. Dr. Nozza's expertise in computing sciences and her commitment to addressing online harms, such as sexism detection, have been invaluable to the success of the workshop.