Francielle Vargas

PhD in Computer Science - Artificial Intelligence

francielleavargas@usp.br

https://franciellevargas.github.io/

Summary

Experienced researcher in Artificial Intelligence and Natural Language Processing with a focus on Language Model Interpretability, Fairness and Hate Speech. 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 with real-world positive societal impact mainly for underrepresented communities in the Global South.

Education

PhD, Computer Science and Computational Mathematics
University of São Paulo (USP)
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 (USP)
2015 - 2017
Thesis: Semantic Clustering of Aspects for Opinion Mining
Artificial Intelligence, Natural Language Processing

BA, Linguistics
Federal University of Minas Gerais (UFMG)
Thesis: An Essay on Socio-Historical Lexicology using Resources from Corpus Linguistics and Computational Linguistics
Computational Linguistics, Corpus Linguistics
2010 - 2014

Advanced Training & Academic Programs

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 Language Processing 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.
2020: Graduate Degree in Software Engineering, PUC Minas, Brazil.
2019: Introduction on the Stars - Astrophysics, University of São Paulo, Brazil.
2016: The R language with an emphasis on probability, ICMC-USP, Brazil.
2015: 4th Summer School in Computer Science, UFMG, Brazil.

Experience

04-2025 to Current: **Postdoctoral Researcher, São Paulo State University, Brazil & Idiap Research Institute, Switzerland.**
The RATIONAL research project is an international and interdisciplinary 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.

01-2024 to 12-2024: **Google LARA PhD Fellowship, Brazil**
I was awarded the Google PhD Fellowship to support my doctoral research in Trustworthy AI, focusing on automated fact-checking and language model interpretability. This fellowship was part of a broader research project coordinated by Dr. Fabrício Benevenuto under the Google Latin America Research Award (LARA).

04-2024 to 04-2024: **Visiting PHD Student, University of Southern California, USA**
Collaborated with researchers at USC under the guidance of Dr. Morteza Dehghani at the Morality and Language Lab. We focused on improving language model interpretability and fairness. 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 invited as a guest researcher to speak at the Conference on Harmful Online Communication. During the event, I shared insights from my research on misinformation, hate speech, and the ethical implications of AI in marginalized communities, contributing to 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 mentoring students on the application of deep learning techniques.

02-2020 to 06-2020: **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, and assisting students with understanding formal models of computation and compiler design concepts.

07-2019 to 12-2023: **Researcher PHD Student, University of São Paulo, Brazil**
Obtained a Ph.D. in 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 misinformation 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 national and international awards.

Awards, Honors and Grants

1. **International Trevisan Prize for Students "AI for Good" 2025, 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 2025, 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 2025 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 2025 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) - PhD Fellowship**
   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. **Diversity and Inclusion Award for EMNLP 2024 and NAACL 2024**
   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 Award with Honorable Mention, Federal University of Minas Gerais 2012 and 2013**

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.

Publications

(*) Equal Contribution

Journal Papers

1. Francielle Vargas, Wolfgang Schmeisser-Nieto, Zohar Rabinovich, Thiago A.S. Pardo, Fabrício Benevenuto. Discourse Annotation Guideline for Low-Resource Languages. In *Natural Language Processing Journal*. Cambridge University Press, pp. 700-743. pdf.
2. Francielle Vargas, Isabelle Carvalho, Thiago Pardo, Fabrício Benevenuto. Context-Aware and Expert Data Resources for Brazilian Portuguese Hate Speech Detection. In *Natural Language Processing Journal*. Cambridge University Press, pp. 435-456. pdf

Conference and Workshop Papers

1. Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, Ameeta Agrawal. *Self-Explaining Hate Speech Detection with Moral Rationales. arXiv cs.CL 2601.03481*. pp. 1-18. pdf
2. Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, Ali Ramezani-Kebrya. *Proceedings of the 40th Annual AAAI Conference on Artificial Intelligence*. pp. 1-15. Singapore, Singapore. pdf
3. Jackson Trager*, Francielle Vargas*, Diego Alves, Matteo Guida, Mikel Nguea-jio, Ameeta Agrawal, Yalda Daryanai, Farzan Karimi-Malekabadi, Flor Miriam Plaza-del-Arco. MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of LLMs through Multi-hop Hate Speech Explanation. *Findings of the Association for Computational Linguistics: EMNLP 2025*. pp. 15709–15740, Suzhou, China. pdf
4. Isadora Salles, Francielle Vargas, Fabrício Benevenuto. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. pp. 6659–6669, Abu Dhabi, UAE. pdf
5. Agostina Calabrese, Christine de Kock, Debora Nozza, Flor Miriam Plaza-del-Arco, Zeerak Talat, Francielle Vargas. Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH). *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*. pp. 1–549, Vienna, Austria. pdf
6. Francielle Vargas, Thiago Pardo, Fabrício Benevenuto. Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection. *Anais do XXXVIII Concurso de Teses e Dissertações (CTD 2025)*, Sociedade Brasileira de Computação. pp. 75–84, Porto Alegre, Brasil. pdf
7. Francielle Vargas, Isadora Salles, Diego Alves, Ameeta Agrawal, Thiago Pardo, Fabrício Benevenuto. Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning. *Proceedings of the 7th Fact Extraction and VERification Workshop (FEVER)*, co-located with EMNLP 2024. pp. 192–204, Miami, United States pdf

8. Francielle Vargas, Samuel Guimarães, Shamsuddeen H. Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo, Fabrício Benevenuto. HausaHate: An Expert Annotated Corpus for Hausa Hate Speech Detection. *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*, co-located with NAACL 2024. pp. 52–58. Mexico City, Mexico. pdf
9. Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, Usman Naseem. Extended Multimodal Hate Speech Event Detection During Russia-Ukraine Crisis - Shared Task at CASE 2024. *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, co-located with EACL 2024 pp. 221–228. St. Julians, Malta. pdf
10. Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago A.S. Pardo, Fabrício Benevenuto (2023). Socially Responsible Hate Speech Detection: Can Classifiers Reflect Social Stereotypes?. *Proceedings of the 14th Recent Advances in Natural Language Processing (RANLP 2023)*. pp. 1187–1196. Varna, Bulgaria. pdf
11. Francielle Vargas, Kokil Jaidka, Thiago A.S. Pardo, Fabrício Benevenuto (2023) Predicting Sentence-Level Factuality of News and Bias of Media Outlets. *Proceedings of the 14th Recent Advances in Natural Language Processing (RANLP 2023)*. pp. 1197–1206. Varna, Bulgaria. pdf
12. Francielle Vargas, Isabelle Carvalho, Wolfgang Schmeisser-Nieto, Fabrício Benevenuto, Thiago A.S. Pardo. NoHateBrazil: A Brazilian Portuguese Text Offensiveness Analysis System. *Proceedings of the 14th Recent Advances in Natural Language Processing (RANLP 2023)*. pp.1180–1186. Varna, Bulgaria. pdf
13. Surendrabikram Thapa, Farhan Jafr, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Le, Usman Naseem. (2023) Multimodal Hate Speech Event Detection CASE Shared Task 4. *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), co-located with EACL 2023*. pp.151-159. Varna, Bulgaria. pdf
14. Francielle Vargas, Isabelle Carvalho, Fabiana R. Góes, Thiago A.S. Pardo, Fabrício Benevenuto. (2022) HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Abusive Language Detection. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. pp. 7174–7183. Marseille, France. pdf
15. Francielle Vargas, Jonas D'Alessandro, Zohar Rabinovich, Fabrício Benevenuto, Thiago A.S. Pardo. (2022) Rhetorical Structure Approach for Online Deception Detection: A Survey. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. pp. 5906-5915. Marseille, France. pdf
16. Francielle Vargas, Thiago A. S. Pardo. Studying Dishonest Intentions in Brazilian Portuguese Texts. *Deceptive AI*. Springer: Communications in Computer and Information Science. vol 1296. pp. 166–178. pdf
17. Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Aaqib Javid, Erdem Yörük. Extended Multilingual Protest News Detection - Shared Task 1, CASE 2021 and 2022.*Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, co-located with EMNLP 2022. pp. 223–228. Abu Dhabi, Arab Emirates. pdf

18. Francielle Vargas, Fabiana R. Góes, Isabelle Carvalho, Fabrício Benevenuto, Thiago A.S. Pardo. Contextual-Lexicon Approach for Abusive Language Detection. *Proceedings of the 13th Recent Advances in Natural Language Processing (RANLP 2021).* pp. 1442-1451. Held Online. pdf
19. Francielle Vargas, Fabrício Benevenuto, Thiago A.S. Pardo (2021). Towards Discourse-Aware Models for Multilingual Fake News Detection. *Proceedings of the Student Research Workshop Associated with RANLP 2021.* pp. 210-218. Held Online. pdf
20. Mateus T. Machado, Thiago A. S. Pardo, Evandro E. S. Ruiz, Ariani Di Felippo, Francielle Vargas. Implicit Opinion Aspect Clues in Portuguese Texts: Analysis and Categorization. *Proceedings of the 15th International Conference on the Computational Processing of Portuguese (PROPOR 2021).* pp. 68-78. Fortaleza, Brazil. pdf
21. Jason R.C. Nurse, Francielle Vargas, Naeemul Hassan, Dakuo Wang, Panagiotis Andriotis, Amira Ghenai, Kokil Jaidka, Eni Mustafaraj, Kenneth Joseph and Brooke Foucault Welles. Towards A Diverse, Inclusive, Accessible and Equitable AAAI International Conference on Web and Social Media (ICWSM). *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM 2021).* pp. 01-12. Held Online. pdf
22. Francielle Vargas, Thiago A. S. Pardo. Linguistic Rules for Fine-Grained Opinion Extraction. *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media (SocialSens), co-located with ICWSM 2020.* pp. 1-6. Held Online. pdf
23. Francielle Vargas, Rodolfo Sanches Saraiva Dos Santos, Pedro Regattieri Rocha. Identifying Fine-Grained Opinion and Classifying Polarity on Coronavirus Pandemic. *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020).* pp. 511-520. Rio Grande, Brazil. pdf
24. Francielle Vargas, Thiago A. S. Pardo. Aspect Clustering Methods for Sentiment Analysis. *Proceedings of the 13th International Conference on the Computational Processing of Portuguese (PROPOR 2018).* pp.365-374. Canela, Brazil. pdf
25. Thiago A. S. Pardo, Jorge Baptista, Magali S. Duran, Maria das Graças Nunes, Fernando Nóbrega, Sandra Aluísio, Ariani Di Felippo, Eloize Seno, Raphael R. Silva, Rafael Anchieta, Henrico Brum, Márcio Dias, Rafael Martins, Erick Maziero, Jackson Souza, Francielle Vargas. The Coreference Annotation of the CSTNews Corpus. *Proceedings of the 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval)*, co-located with SEPLN 2017. pp. 102-112. Murcia, Spain. pdf

Non-peer reviewed articles

1. Francielle Vargas, and Thiago A. S. Pardo. Empirical Study on Clustering and Hierarchical Organization of Aspects for Opinion Mining. *Technical Institute of Mathematics and Computer Science's Reports, University of São Paulo.* São Carlos, Brazil. no.418. 48 pp. 2017.pdf

---

Talks & Presentation

2026: **Roundtable Lead at the Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA) Workshop, International Conference on Learning Representations (ICLR 2026), Rio de Janeiro, Brazil**. Roundtable at the AFAA Workshop, co-located with ICLR 2026, I will lead a roundtable discussion focused on algorithmic fairness across alignment procedures and agentic

systems. The session will bring researchers to examine how alignment strategies, learning objectives, and agentic behaviors interact with fairness considerations, highlighting open challenges in responsible and trustworthy machine learning.

2025: **Poster Presentation at the Applied Social Media Lab Synthesizer & Open Showcase, Berkman Klein Center for Internet & Society at Harvard University, Boston, United States**.
Posters: *Brazil#WithoutHate: Self-Explaining and Moral-Aware AI for Hate Speech Detection* and *Factuality and Transparency Are All RAG Needs! Self-Explaining Contrastive Evidence Re-ranking for Retrieval-Augmented Generation.*
During the Applied Social Media Lab Synthesizer & Open Showcase at the Berkman Klein Center, I presented two posters introducing my research on self-explaining methods to ensure factuality and transparency for trustworthy RAG and hate speech detection systems.

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 Dr. Morteza Dehghani to talk at the 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.

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 2021, 2022, 2023**: I served as co-chair of the

Datasets Track, and as Diversity, Inclusion, & 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. **Area Chairing**: I have served as an Area Chair for leading Artificial Intelligence conferences, including:
   - International Joint Conference on Neural Networks (IJCNN)
2. **Journal Reviewer**: I am a reviewer for some of the most prestigious international journals in Natural Language Processing (NLP), including:
   - Natural Language Processing.
   - Language Resources and Evaluation.
   - Online Social Networks and Media.
   - Expert Systems with Applications
3. **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)

Conference Participation / Presentation

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

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 (HateBRXplain). This resource 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.

Benchmark Datasets

HateBRXplain and HateBRMoralXplain: Benchmark Datasets with Hate Speech and Moral Human-Annotated Rationales for Explainable HS Detection.
MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of Large Language Models
HateBR: A Benchmark Dataset for Explainable HS Detection in BR Portuguese.
FactNews: A Benchmark Dataset for Sentence-Level Factuality Prediction.
HausaHate: A Benchmark Expert Dataset for Hausa HS 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 HS Detection.
B+M: Contextual BoW with Feature Saliency for Explainable HS Detection.
SSA: Post-Hoc Counter-Stereotype Explanations for Bias Assessment in HS Classifiers.

Softwares

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

Technical Skills

Python, NLTK, spaCy, Keras, Tensorflow, PyTorch, SQL, RDF(s), OWL, Unix.

Languages

English: Professional. **Spanish**: Basic. **Portuguese**: Native.

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 papers published to EMNLP 2024 and 2025. Our research addresses key topics such as explainability and interpretability, bias mitigation, 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 Explainable Deep Neural Networks for Responsible AI (DeepXplain), co-located with International Joint Conference on Neural Networks 2025 in Italy. The workshop aims to foster advancements in interpretability 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. Debora Nozza in person at EMNLP 2025 in Miami. We later worked together in the ACL community as members of the organizing committee of the 9th Workshop on Online Abuse and Harms (WOAH), co-located with the ACL 2025, which she co-chaired. Our interaction was focused on workshop organization and program-related activities, reflecting a shared interest in hate speech research and socially-aware and responsible NLP. I am currently collaborating with her postdoctoral researcher, Dr. Arianna Muti, on a project I am leading focused on language model interpretability, hate speech, and multicultural contexts. Debora has also been exceptionally generous and 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. Fabrício 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. As a result of our collaboration, we have published 12 papers in top-tier NLP conferences, and I also had the opportunity to co-advise his master's student, which resulted in an international publication at COLING 2025.

Dr. Morteza Dehghani. Director of the Center for Computational Language Sciences and Professor of Psychology and Computer Science at 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 published 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.