Daniel Scalena

+39 3** *** ***8
danielsc4.it
Milan, Italy
daniel-scalena
scalena99@gmail.com
danielsc4

Education

University of Milano - Bicocca | University of Groningen, CLCG *Milan, Italy* **|** *Groningen, Netherlands*
Double Doctorate (Ph.D) in Computer Science - NLP *Nov 2023 – Present*

Research focuses on the use of interpretability as a tool to make generative models safer, more reliable and less toxic in order to extend and improve their real-world applications

University of Milano - Bicocca *Milan, Italy*
Master of Science (M.Sc) in Computer Science *Oct 2021 – Oct 2023*

Thesis: On the explainability of Large Language Models detoxification
Curriculum: AI & Machine Learning
Final Grade: 110/110, with honors

University of Milano - Bicocca *Milan, Italy*
Bachelor of Science (B.Sc) in Computer Science *Oct 2018 – Jul 2021*

Thesis: Hate speech detection on social networks using state-of-the-art Natural Language Processing techniques
Final grade: 110/110, with honors

Work Experience

University of Groningen, Center for Language and Cognition *Apr 2023 – Jul 2023*
Research Intern *Groningen, Netherlands*

- Study, research and development of RLHF/RLAIF and fine-tune algorithms applied to generative language models to modify their behaviour in generating hate speech, effectively automating and controlling the detoxification process and counter-narrative generation;
- Research on the interpretability of the models themselves, analyzing their shift as a result of the post-training procedures adopted.

University of Milano - Bicocca, MIND Lab *Jun 2022 – Apr 2023*
Assistant Researcher *Milan, Italy*

- Research project commissioned by the Italian Ministry of Justice and CINI (Italian Interuniversity Consortium for Computer Science).
- Open Relation Extraction on criminal sentences - Improving state-of-the-art for technical and legal texts understanding using text mining techniques, seq2seq models and transformer-based Large Language Models.

TESTUDO *Jun 2022 – Aug 2022*
Machine Learning Engineer *Milan, Italy*

- Worked with University MIND Lab team developing deep learning models to predict timestamps regarding working hours;
- I had the opportunity to explore and train deep neural networks on complex and non-ideal data, improving the productivity of the company's automatic systems.

University of Milano - Bicocca *Mar 2021 – May 2021*
Intern *Milan, Italy*

- University internship in the field of Hate Speech detection using NLP techniques on Tik-Tok social network;
- Achieved excellent performance thanks to the use of lexicons and large language models with transformers-based architecture.

Publications

🔗 **Let the Models Respond: Interpreting Language Model Detoxification Through the Lens of Prompt Dependence**
Daniel Scalena, Gabriele Sarti, Malvina Nissim, Elisabetta Fersini
Extended abstract at the Sixth BlackboxNLP Workshop (EMNLP 2023)

🔗 **MIND at SemEval-2023 Task 11: From Uncertain Predictions to Subjective Disagreement**
Giulia Rizzi, Alessandro Astorino, **Daniel Scalena**, Paolo Rosso, Elisabetta Fersini
Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)

Projects, Courseworks and Research works

Let the Models Respond: Interpreting the Detoxification process of LMs *Spring 2023*
Ongoing research project *Journal*

- Reached enhancement and detoxification of Generative Language Models;
- Aim not to limit model generation with toxic prompts but to foster counter-narrative.
- LMs interpretability techniques for detoxification process evaluation after fine-tuning and/or RLHF;

Reward LM *Spring 2023*
Open source Python library *GitHub*

- Python toolkit library enabling Fine-Tuning and Reinforcement Learning from *AI* Feedback (RLAIF) of Generative Large Language Models;
- Deeply integrated with HuggingFace and Accelerate frameworks;
- A structured methodology for measuring model toxicity is provided, allowing measurement with datasets recognised in the scientific literature.

Shared pre-trained detoxified Language Models *Summer 2023*
Open source contribution *HuggingFace*

- Sharing a series of models trained to be detoxified, following different techniques listed in a publication in progress;
- Detoxification process reached up to -30% on SOTA benchmark dataset;
- The models, in terms of configuration and number of parameters, currently represent the state of the art on the HuggingFace leaderboard.

From Uncertain Predictions to Subjective Disagreement *Spring 2023*
Research project & pubblication *ACL Anthology*

- Participation of the research lab MIND from UniMiB in the SemEval 2023 task related to Learning With Disagreements (Le-Wi-Di);
- Study the identification of the level annotator's agreement/disagreement;
- Explored the correlation between disagreement and uncertainty that a model, based on several linguistic characteristics, could have on the prediction of a given gold label.

-> Other projects and Open Source contributions *are available on my personal GitHub profile*

Language Skills

Italian	Native Language
English	Badge Bbetween Foreign Languages – English C1 🔗
French	B1

Activities

2023 European Researchers Night	GroNLP collaborator, building demo tools on LMs Interpretability