

Education

University of Milano - Bicocca

Master of Science (M.Sc) in Computer Science

Thesis: Title TBD

Curriculum: AI & Machine Learning

Current Grade: 29.7/30 (weighted average)

Milan, Italy

Oct 2021 – Oct 2023

University of Milano - Bicocca

Bachelor of Science (B.Sc) in Computer Science

Thesis: Hate speech detection on social networks using state-of-the-art Natural Language Processing techniques

Final grade: 110/110 cum Laude

Milan, Italy

Oct 2018 – Jul 2021

Work Experience

University of Groningen, Center for Language and Cognition

Research Intern

Apr 2023 – Jul 2023

Groningen, Netherlands

- Study, research and development of Reinforcement Learning algorithms applied to generative models (LLMs) to modify their behaviour in generating hate speech, effectively automating and controlling the detoxification process;
- Collaborated further on the interpretability of the models themselves, researching the motivations of these generations.

University of Milano - Bicocca, MIND Lab

Assistant Researcher

Jun 2022 – Apr 2023

Milan, Italy

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

TESTUDO

Machine Learning Engineer

Jun 2022 – Aug 2022

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

Intern

Mar 2021 – May 2021

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 and thesis

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)

2023

Publication

Natural language processing techniques for recognition of hate speech on social networks

Daniel Scalena, Elisabetta Fersini

2021

Thesis

Projects, Courseworks and Research works

Let the Models Respond: Interpreting the Detoxification process of LMs

Ongoing research project

Spring 2023

[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

Open source Python library

Spring 2023

[GitHub](#)

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

Shared pre-trained detoxified Language Models

Open source contribution

Summer 2023

[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

Research project & publication

Spring 2023

[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.

Music Genre classification with NLP

Coursework

Winter 2022

[GitHub](#)

- Led the development of the project. Two models (SVM and Neural Networks) are compared in music genre classification where NLP techniques are proposed and evaluated to increase performance;
- Achieved and improved top performance on the Kaggle challenge leaderboard.

-> **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 DELF, B1

Activities

2023 European Researchers Night

GroNLP collaborator, building demo tools on LMs Interpretability