

PERSONAL INFORMATION

Giovanni Puccetti



 Vicolo degli Orefici 4, Siena.

 +39 351 9443040

 giovanni.puccetti@sns.it / g.puccetti92@gmail.com

 <https://gpucce.github.io>

 <https://github.com/gpucce>

Gender Male | Date of birth 15/08/1992 | Nationality Italian

SUMMARY

I have studied Mathematics full-time up to a Masters degree, I chose this course of studies because I enjoy engaging in analytical thinking and problem solving activities. After my master degree I moved towards more computation involved studies by pursuing a PhD in Data Science. In the PhD program where I am currently enrolled, I investigate applications of statistics and deep learning to Natural Language Processing.

I have mainly pursued two research paths, on one side I focused on information extraction from patents by identifying technical entities such as users, technologies, problems and solutions from patents corpora. On the other side I have worked on widening the understanding of large language models from a linguistic and mechanistic point of view. To do so I have participated in projects involving the training of language models from scratch and assessing their performance on benchmarks widely used in the NLP literature.

My extra curricular activities so far included, basketball, travelling and language studies (Italian, English and Dutch).

PROFESSIONAL EXPERIENCE

- 2019 (three months) Research fellowship University of Pisa (Paid).
Prot. 668 del 12.02.19 Rep.59.
- 2018 (three months) Teaching assistant for the course Experimental Design and Data Analysis, Vrije University. (Paid)
Contract N. HRM201800034lh
- 2017 (three months) Teaching assistant for the course Statistics Simulation and Optimization, University of Amsterdam. (Paid)
Contract N. PZ-2017/CoS/6134a

EDUCATION

- 03/10/2022-21/12/2022 Period abroad: RIKEN, High Performance Artificial Intelligence Systems Research Team, Tokyo, Japan. Working with Aleksandr Drozd.
Theme: Training of Transformer-based Image-Text Alignment Models
- 01/02/2022-31/05/2022 Period abroad: University of Copenhagen, Copenhagen Center for Social Data Science (SO-DAS), Copenhagen, Denmark. Working with Anna Rogers.
Theme: Improving the mechanistic understanding of transformer-based language models
- 01/11/2019-currently Attending PhD programme in Data Science at Scuola Normale Superiore di Pisa
Thesis prospective title: Transformers: from a Computational Linguistics Perspective to Applications in Technical Documentation
- 01/11/2019 PhD Data Science. Supervisors: Felice Dell'Orletta, Gualtiero Fantoni, Fosca Giannotti.
- 14/11/2018 Master degree in Mathematics. (7.99/10)
Thesis Title: Average Nearest Neighbor Degree in Geometric Inhomogeneous Random Graph

	date 14/11/2018. University of Amsterdam
2016/2017-2018/2019	Attended the Master in Mathematics at University of Amsterdam.
11/03/2016	Bachelor degree in Mathematics. (105/110) <i>Thesis Title:</i> Calcolo del Terzo Gruppo di Omotopia della Due-Sfera Diploma n. 20152966071, date 11/03/2016. University of Pisa
2011/2012-2014/2015	Attended the course in Mathematics at University of Pisa.
10/07/2011	Maturità classica. (86/100)
2006-2011	Attended High School “Liceo classico Enea Silvio Piccolomini” of Siena. Lasting my studies I attended different courses of language learning: Malta (English Language Academy), Galles, Canada (English as a second language). Paris (Alliance Francaise), Bruxelles.

LANGUAGES

Mother tongue Italian

Other languages	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
French	B2	B2	B2	B2	B2
Dutch	B1	B1	B1	B1	B1
English	C1	C1	C1	C1	C2

Levels: A1 and A2: Basic user – B1 and B2: Independent user – C1 and C2: Proficient user
[Common European Framework of Reference for Languages](#)

IT skills **Python language.** Main packages: pytorch, pytorch-lightning, huggingface-transformers, huggingface-datasets, scikit-learn, spacy, nltk.
Julia language. Main packages: Jarvis.jl, Pluto.jl, DataFrames.jl, Statistics.jl, GLM.jl, MLJ.jl.
The use of scientific computing software, moderate ability in Matlab and R.

Software Software for the study of outlier dimensions: <https://github.com/gpucce/outliersvsfreq>
Open source contributor: <https://github.com/gpucce/>
Participate as a mentor in 2022 Google Summer of Code

Event Organization Participated in organizing the webinar series **ENG4GED** <http://b4ds.unipi.it/eng4ged/>
Organized the event **A Glimpse of Complexity the Lens of AI** <https://www.eventbrite.it/e/biglietti-a-glimpse-of-complexity-the-lens-of-ai-83581621881>

Extra Curricular activities Blog explaining mathematics through animations: <https://gpucce.github.io/gifblogging>
Sports: Basketball, Rugby.
From 2000 to 2008 I was a boy scout in Siena.

PUBLICATIONS

- [1] **Giovanni Puccetti**, Filippo Chiarello, and Gualtiero Fantoni. “A simple and fast method for Named Entity context extraction from patents”. In: *Expert Systems with Applications* 184 (2021), p. 115570. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2021.115570>. URL: <https://www.sciencedirect.com/science/article/pii/S0957417421009751>.

- [2] **Giovanni Puccetti**, Alessio Miaschi, and Felice Dell’Orletta. “How Do BERT Embeddings Organize Linguistic Knowledge?” In: *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Online: Association for Computational Linguistics, June 2021, pp. 48–57. DOI: 10.18653/v1/2021.deelio-1.6. URL: <https://aclanthology.org/2021.deelio-1.6>.
- [3] **Giovanni Puccetti** et al. “B4DS @ PRELEARN: Ensemble Method for Prerequisite Learning (short paper)”. In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*. Ed. by Valerio Basile et al. Vol. 2765. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: <http://ceur-ws.org/Vol-2765/paper107.pdf>.
- [4] **Giovanni Puccetti** et al. “Outliers Dimensions that Disrupt Transformers Are Driven by Frequency”. In: *Findings of EMNLP, 2022* (to appear). arXiv, 2022. DOI: 10.48550/ARXIV.2205.11380. URL: <https://arxiv.org/abs/2205.11380>.
- [5] **Giovanni Puccetti** et al. “Technology identification from patent texts: A novel named entity recognition method”. In: *Technological Forecasting and Social Change* 186 (2023), p. 122160. ISSN: 0040-1625. DOI: <https://doi.org/10.1016/j.techfore.2022.122160>. URL: <https://www.sciencedirect.com/science/article/pii/S0040162522006813>.

PREPRINTS

Vito Giordano, **Giovanni Puccetti**, Filippo Chiarello, Tommaso Pavanello, Fantoni Gualtierio. "Unveiling the Inventive Process from Patents by Extracting Problems, Solutions and Advantages with Natural Language Processing". Link: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4223458