Martin Tutek, PhD

https://mttk.github.io/
martin.tutek at gmail.com

Research Interests

Natural Language Processing, Interpretability, Reasoning in Large Language Models

Education

Ph.D. In Computer Science, Faculty of Electrical Engineering and Computing, University of Zagreb — **Apr 2016 – July 2022**

- Thesis: "*Extending the recurrent neural network model for improved compositional modeling of text sequences*"; advisor: prof. dr. sc Jan Šnajder

M.Sc. In Computer Science, Faculty of Electrical Engineering and Computing, University of Zagreb — 2012 – 2014

B.Sc. In Computer Science, Faculty of Electrical Engineering and Computing, University of Zagreb — 2009 – 2012

Professional Experience

Postdoctoral Researcher, Technion — **Feb 2024 - Present**

- *Mechanistic interpretability*

Postdoctoral Researcher, UKP Lab, Technische Universität Darmstadt — **Sep 2022 - Dec 2023**

- *Learning structure augmented representations of long textual documents, improving LLMs through training data augmentation and transformation*
- *Teaching: Deep Learning for NLP (lecturer)*

Research Assistant, TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb — **Feb 2016 – Aug 2022**

- Teaching: *Artificial Intelligence (Head TA; lab assignments; lectures; course material; 2016–2022)*, *Deep Learning (lab assignments; lectures; course material; 2017–2022)*, *Text Analysis and Retrieval (lectures; course material; 2017–2022)*
- *Supervision of 10+ MA students and 10+ BA students (co-mentor: Jan Šnajder)*

Consultant, European Commission, Joint Research Centre, Ispra, Italy — **May 2015 – Sep 2015**

Trainee, European Commission, Joint Research Centre, Ispra, Italy — **Sep 2014 – Feb 2015**

- Applied NLP for updating terminology of the *Sendai Framework for Disaster Risk Reduction*

Service

- **Area Chair:** Interpretability and Analysis of Models for NLP
  ACL 2023, EMNLP 2023, ARR Dec 2023 - Present
- **Conference Reviewer**
  ARR Nov *2021 – Oct 2023 (outstanding reviewer Oct 2023)*
  EMNLP *2018 – 2022*
  ACL *2018 – 2022*
  COLM 2024
- **Journal Reviewer**
  Automatika 2020, 2021
  Artificial Intelligence 2021, 2022

- **Summer School Lecturer**
  Intl' Summer School of Data Science in Split, practical sessions - *Random Forests and Gradient Boosting* (2016); *Generative Adversarial Networks* (2017)

Selected Publications

- Puerto, H., **Tutek, M.**, Aditya, S., Zhu, X., & Gurevych, I. (2024). Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs. Arxiv preprint.
- Jelenić, F., Jukić, J., **Tutek, M.**, Puljiz, M., & Šnajder, J. (2024). Out-of-Distribution Detection by Leveraging Between-Layer Transformation Smoothness. ICLR 2024.
- Sachdeva, R., **Tutek, M.**, & Gurevych, I. (2024). CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration. EACL 2024.
- Jukić, J., **Tutek, M.**, & Šnajder, J. (2023). Easy to Decide, Hard to Agree: Reducing Disagreements Between Saliency Methods. Findings of the Association for Computational Linguistics: ACL 2023
- **Tutek, M**., & Snajder, J. (2022). Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability. IEEE Access.
- Obadić, L., **Tutek, M.**, & Šnajder, J. (2022). NLPOP: a Dataset for Popularity Prediction of Promoted NLP Research on Twitter. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 286-292).
- **Tutek, M.** & Šnajder, J. (2020). Staying True to Your Word:(How) Can Attention Become Explanation?. In Proceedings of the 5th Workshop on Representation Learning for NLP (pp. 131-142).
- **Tutek, M.** & Šnajder, J. (2018). Iterative Recursive Attention Model for Interpretable Sequence Classification. In Proceedings of the 2018 EMNLP Workshop: Analyzing and interpreting neural networks for NLP.
- **Tutek, M.**, Glavas, G., Šnajder, J., Milić-Frayling, N., & Dalbelo Basic, B. (2016, October). *Detecting and Ranking Conceptual Links between Texts Using a Knowledge Base*. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 2077-2080).
- **Tutek, M.**, Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D. and Šnajder, J. (2016). Takelab at semeval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 464-468).