LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Evaluating Faithfulness of Post-hoc Explanation Methods for Transformer-based Text Classification**

Supervisor: Yuetian Lu

Examiner: Prof. Dr. Hinrich Schütze

Level: BSc

Summary: Modern Transformer-based text classifiers achieve strong performance, but their decisions are often difficult to interpret. This thesis will systematically compare widely used *post-hoc local explanation* methods for Transformer classifiers, including gradient-based saliency, Integrated Gradients, attention-based variants (e.g., attention rollout / attention flow), and perturbation-based token occlusion. The student will evaluate these methods on one or two standard text classification tasks (e.g., sentiment analysis and/or toxicity detection) using faithfulness metrics such as deletion/insertion curves and (when explanations are converted into rationales) sufficiency/comprehensiveness. A small-scale human plausibility check can be added as a sanity check of whether highlighted evidence aligns with human intuition. The expected outcome is a reproducible evaluation pipeline and practical recommendations on when different explanation methods are reliable.

Requirements: Solid programming skills in Python; basic knowledge of machine learning and NLP; familiarity with (or willingness to learn) PyTorch and HuggingFace Transformers; careful experimental practice (reproducibility, ablations) and ability to write a clear thesis in English.

References:

- Mukund Sundararajan, Ankur Taly, and Qiqi Yan (Aug. 6–11, 2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html
- Vitali Petsiuk, Abir Das, and Kate Saenko (2018). "RISE: Randomized Input Sampling for Explanation of Black-box Models". In: *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, p. 151. URL: http://bmvc2018.org/contents/papers/1064.pdf
- Sarthak Jain and Byron C. Wallace (June 2019). "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://aclanthology.org/N19-1357/
- Samira Abnar and Willem Zuidema (July 2020). "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://aclanthology.org/2020.acl-main.385/
- Jay DeYoung et al. (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408/