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

Thesis proposal

Topic: **Warm-Starting Active Learning with Synthetic Data for Imbalanced Hate Speech Classification**

Supervisor: Ahmad Dawar Hakimi

Examiner: Prof. Hinrich Schütze

Level: Bsc

Summary: Hate speech and online toxicity detection are classic examples of highly imbalanced classification problems: harmful content is much rarer than benign text, but far more important to detect reliably. Standard supervised learning struggles in this setting, models are biased toward the majority class, recall on harmful content is low, and collecting enough labeled toxic examples is costly. Recent large language models (LLMs) make it feasible to generate labeled synthetic hate speech, borderline cases, and non-toxic examples, which can be used to "warm-start" smaller classifiers before running active learning (AL) on real, unlabeled comments. In this project, the student will investigate whether LLM-generated synthetic data targeted at minority (harmful) classes can improve downstream active learning on a real-world hate speech or toxicity dataset with strong class imbalance.

Research questions:

1. How does pretraining a hate speech classifier on LLM-generated synthetic data affect subsequent active learning performance compared to starting from a small real seed set, under the same labeling budget?
2. How does synthetic warm-starting influence the behavior of common acquisition strategies (e.g., uncertainty sampling, diversity-based sampling) in imbalanced settings, in terms of which examples are selected and how quickly minority-class performance improves?
3. Under what conditions (prompt style, amount of synthetic data, LLM choice, subtype coverage) does synthetic warm-starting degrade downstream active learning performance or distort the decision boundary compared to training only on real data?

Supervision can be provided in either German or English.

Requirements:

- Strong interest in NLP, text classification, and hate speech / toxicity detection.
- Solid Python skills and experience with deep learning frameworks (e.g., PyTorch, Hugging Face, Transformers).
- Familiarity with Encoder Models and LLMs.
- Ideally, some prior exposure to active learning or class imbalance methods (e.g., oversampling, class weighting), but this can be learned during the thesis.

References:

- Xinyi Gao et al. (2025). "A comprehensive survey on imbalanced data learning". In: *arXiv preprint arXiv:2502.08960*
- Xuanli He et al. (2022). "Generate, annotate, and learn: NLP with synthetic text". In: *Transactions of the Association for Computational Linguistics* 10, pp. 826–842
- Mihai Nadăș, Laura Dioșan, and Andreea Tomescu (2025). "Synthetic data generation using large language models: Advances in text and code". In: *IEEE Access*

- Ricardo Barata et al. (2021). "Active learning for imbalanced data under cold start". In: *Proceedings of the Second ACM International Conference on AI in Finance*, pp. 1–9
- Mateusz Bystroński et al. (2025). "SMOTExT: SMOTE meets Large Language Models". In: *arXiv preprint arXiv:2505.13434*