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

Thesis proposal

Topic: **Hate speech detection in a low-resource setting**

Supervisor: Axel Wisiorek, Haotian Ye

Examiner: Hinrich Schütze

Level: BSc / MSc

Summary: Large language models (LLMs) are proving to be extremely useful, as they can perform NLP tasks without requiring labeled data, which either do not exist or are very limited for a large portion of the world's languages. These languages are often considered low-resource in the field of NLP due to the scarcity of annotated datasets.
However, these models, which are trained on massive amounts of online data and represent multilingual capabilities, show a significant imbalance in their training data in favor of English. This imbalance raises concerns about their performance for low-resource languages, particularly in nuanced tasks such as detecting hate speech or analyzing sentiment.
The goal of this thesis is to investigate whether LLMs can be effectively applied to NLP tasks for low-resource languages, with a particular focus on hate speech detection or sentiment analysis. This will involve creating a labeled dataset in a low-resource language (with respect to hate speech or sentiment analysis) and using this dataset to evaluate the performance of LLMs.
Possible directions and experiments include:

- Annotate and validate a dataset for hate speech detection or sentiment analysis in a low-resource language using the existing data collection pipeline.
- Evaluate the performance of LLMs (e.g., GPT models) on hate speech detection or sentiment analysis using the newly annotated dataset.
- Analyze the strengths and weaknesses of zero-shot and few-shot learning approaches on the task, particularly for low-resource settings.
- Test and provide feedback on a client-side tool for NLP tasks (e.g., hate speech detection or sentiment analysis) and assess its usability and performance for low-resource languages.
- Conduct experiments to compare the performance of LLMs with other baseline approaches, including smaller or specialized models trained on low-resource languages.

Requirements: Good programming and data processing skills (preferably using Python), enthusiasm for multilingual language processing, and proficiency in a language considered low-resource for hate speech or sentiment analysis (to be discussed with the supervisors).

References:

[1] Paul Röttger, Debora Nozza, et al. (Dec. 2022). "Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5674–5691. URL: https://aclanthology.org/2022.emnlp-main.383

[2] Antonis Maronikolakis et al. (May 2022). "Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 1089–1104. DOI: 10.18653/v1/2022.findings-acl.87. URL: https://aclanthology.org/2022.findings-acl.87

[3] Paul Röttger, Haitham Seelawi, et al. (July 2022). "Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models". In: *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Seattle, Washington (Hybrid): Association for Computational Linguistics, pp. 154–169. DOI: 10.18653/v1/2022.woah-1.15. URL: https://aclanthology.org/2022.woah-1.15

[4] Janis Goldzycher et al. (July 2023). "Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data". In: *The 7th Workshop on Online Abuse and Harms (WOAH)*. Toronto, Canada: Association for Computational Linguistics, pp. 187–201. URL: https://aclanthology.org/2023.woah-1.19