LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Pragmatic categories in hate speech annotation**

Supervisor: Axel Wisiorek

Examiner: Axel Wisiorek

Level: BSc/MSc

Keywords: hate speech, datasets, annotation, pragmatics, discourse analysis

Summary: Hate speech is not only characterized by offensive vocabulary, but often by pragmatic and discourse-level strategies such as irony, provocation, dehumanisation, or stance-taking. Many existing hate speech datasets focus on surface-level labels (e.g., hate vs. non-hate) and neglect these communicative aspects. As a result, current models often struggle with implicit or context-dependent forms of hateful discourse.

The goal of this thesis is to explore whether **pragmatic annotation categories** can be reliably applied by non-expert annotators and whether such annotations provide additional value for hate speech analysis. The thesis will focus on a small, well-defined set of pragmatic categories (e.g., speech acts, framing, stance) and apply them to a subset of an existing hate speech dataset.

Students will annotate pragmatic categories such as speech acts, stance, framing, irony, provocation, and dehumanisation, guided by conceptual categories inspired by HateCheck.

Possible directions and experiments include:

- Design or adapt a concise annotation scheme for selected pragmatic categories relevant to hate speech (e.g., speech acts, framing, stance-taking).
- Apply the annotation scheme to a subset of an existing hate speech dataset and document annotation challenges.
- Measure inter-annotator agreement to assess the reliability and clarity of the pragmatic categories.
- Perform an exploratory qualitative analysis showing how pragmatic annotations capture implicit or context-dependent hate beyond surface-level keywords.

Possible M.Sc. extensions:

- Compare pragmatic annotation schemes or theoretical frameworks.
- Investigate correlations between pragmatic categories and model performance or error patterns.
- Explore semi-automatic or model-assisted annotation.
- Conduct a small-scale classification or probing experiment using pragmatic labels.

Requirements: Basic programming and data handling skills in Python; interest in linguistics/pragmatics; willingness to work with annotated text data.

For MSc students, familiarity with NLP methods, corpus linguistics, or experimental design is desirable. Prior experience with annotation studies, machine learning, or discourse analysis can be incorporated depending on the student's background.

Resources: Existing hate speech datasets (e.g., Davidson et al. 2017, Waseem & Hovy 2016, HateCheck), annotation guidelines and tools provided by the supervisor.

Core References:

- Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*. In Proceedings of the International AAAI Conference on Web and Social Media. `https://doi.org/10.1609/icwsm.v11i1.14955`
- Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter*. In Proceedings of the NAACL Student Research Workshop (pp. 88–93). San Diego, California: ACL. `https://aclanthology.org/N16-2013/`
- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). *HateCheck: Functional tests for hate speech detection models*. In Proceedings of the 59th Annual Meeting of the ACL

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 41–58. ACL. https://aclanthology.org/2021.acl-long.4/

Supplementary References:

- Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). *Understanding abuse: A typology of abusive language detection subtasks*. In Proceedings of the First Workshop on Abusive Language Online (pp. 78–84). https://aclanthology.org/W17-3012/
- Vidgen, B., & Derczynski, L. (2020). *Directions in abusive language training data: Garbage in, garbage out*. PLOS ONE, 15(12). https://arxiv.org/abs/2004.01670
- Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). *Toxicity detection: Does context really matter?* In Proceedings of the 58th Annual Meeting of the ACL (pp. 4296–4305). https://aclanthology.org/2020.acl-main.396/
- Culpeper, J. (2011). *Impoliteness: Using language to cause offence*. Cambridge University Press.
- Fortuna, P., & Nunes, S. (2018). *A survey on automatic detection of hate speech in text*. ACM Computing Surveys, 51(4). https://dl.acm.org/doi/10.1145/3232676