Universität Zürich[UZH]

URPP Digital Religion(s)

Link to Publication DB

P8: Argument Structures in the Automatic Detection of Intolerance and Extremism

P8 aims to detect hate speech and associations to religion, spirituality and related topics.

We improve hate speech detection, particularly develop robust systems using large language models (LLMs). We visualize associations to religion, spirituality and faith.

We offer content analysis to all URPP projects.

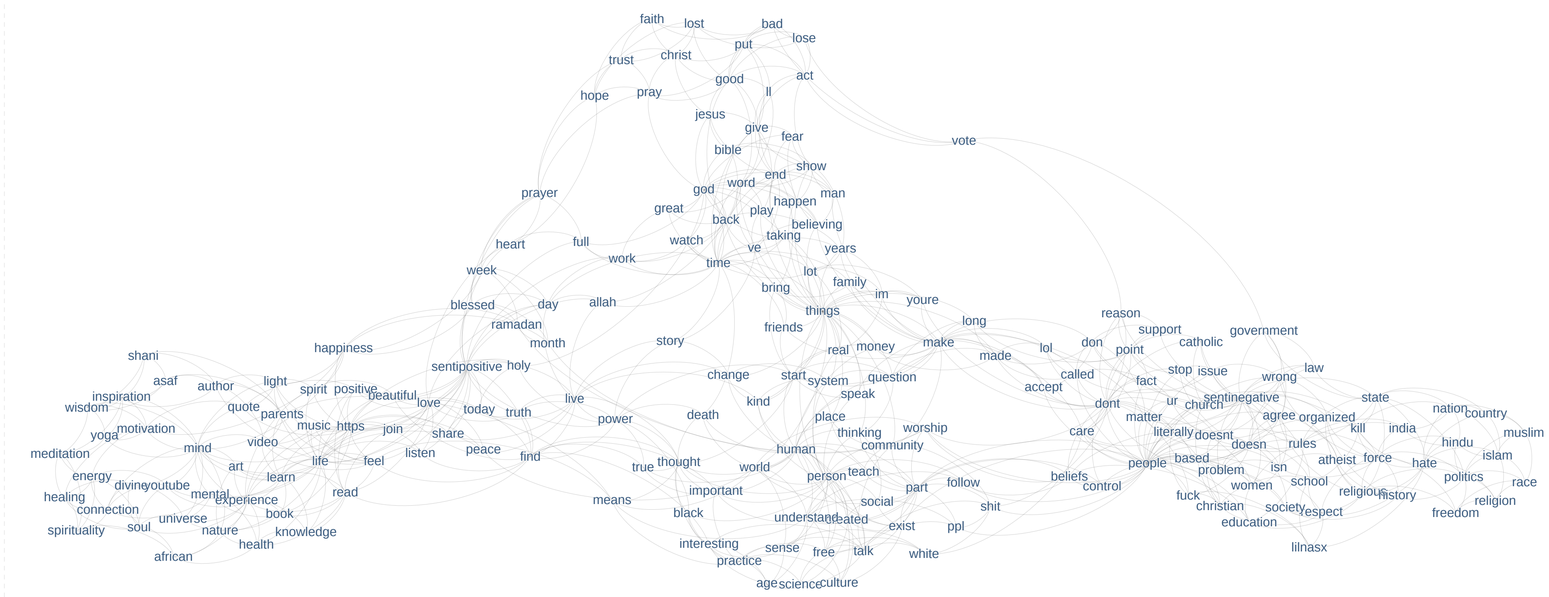

Fig.1: Associations to *sprirituality* (left), *religion* (right) and *faith* (top) generated automatically from tweets. Obs. the added pseudowords *sentipositive/negative*

1) Prompting LLMs to improve Hate Speech Detection (HSD)

Objective
Improve the detection of hate speech in social media, using the latest AI methods.

Methods
We use the currently best methods such as LLMs, or similar transformer-based models. Prompting LLMs, for instance, to make sure that the object of hate is a protected group, to detect jokes etc.

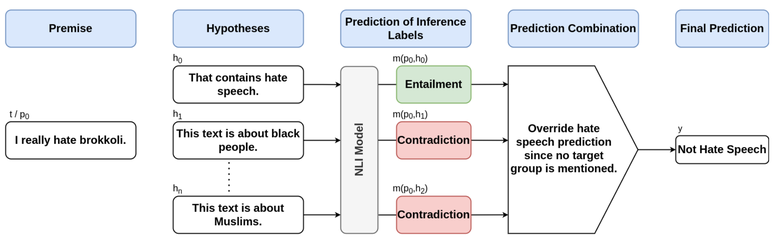

Fig. 2. Standard zero-shot entailment predictions would wrongly predict the input text as containing hate speech. With additional hypotheses, it is possible to check if a protected group is targeted and to override the original prediction.

We use the following strategies:
- Filter by groups: e.g. “This text is about Muslims.”
- Filter by characteristics: e.g. “This text is about race.”
- Filter counterspeech: e.g. “This text supports [X].”
- Filter reclaimed slurs: e.g. “This text is about myself.”
- Catch dehumanization: e.g. “This text is about rats.”

Results
These strategies improve HSD (by up to 10%, from 69.6 to 79.6%) , and offer intuitive ways also for non-programmers to improve performance or adapt it to specific domains and tasks. We have several publications at international conferences (Goldzycher and Schneider 2022, Goldzycher et al. 2023).

Our Contributions to the URPP

- Hate Speech impacts many religious communities. Its detection is central to their well-being & safety
- Additional financing by the Swiss Government Agency of Communication (BAKOM)
- Associations to religion, faith and spirituality helps to interpret their roles in society (with P11)
- Religion Monitor (with P12)
- Coursebook enabling Media Content Analysis (Schneider 2024)
- Detection and analysis of religious metaphors together with Johannes Fröh (Schneider fc.)

2) Increasing robustness of HSD with adversarial attacks

Objective
Improve the detection of hate speech in social media, based on a large systematic collection of difficult instances. We introduce GAHD (Goldzycher et al. 2024), a new German Adversarial Hate speech Dataset of 11,000 examples.

Methods
Datasets sourced from social media suffer from systematic gaps and biases. Adversarial datasets, collected by exploiting model weaknesses, can fix this problem. They are created by tasking annotators to trick the model. We explore new strategies to support annotators in tricking the model, as shown in Figure 3. Figure 4 demonstrates the impact of these strategies on model performance.

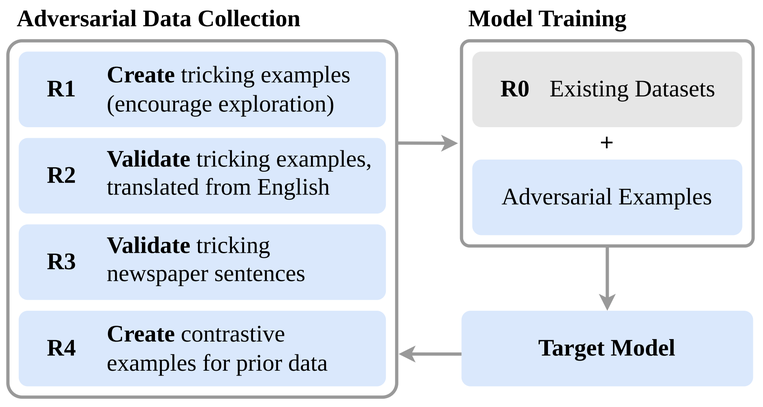

Fig.3: Overview of our 4 adversarial attack strategies

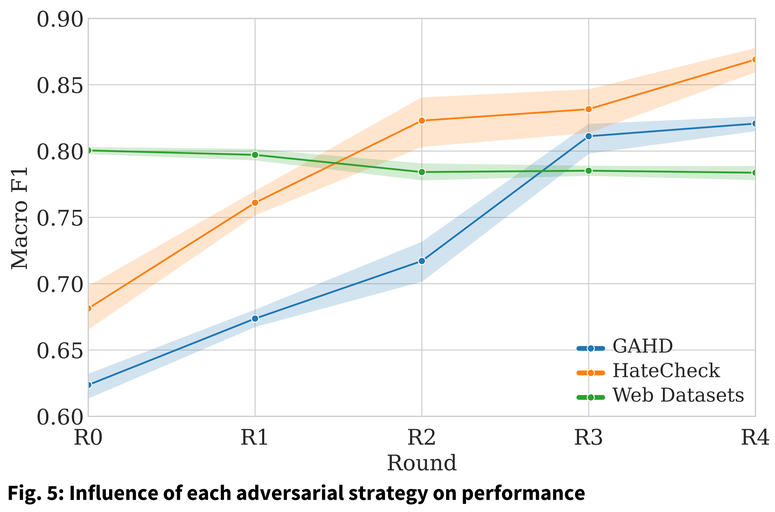

Fig. 5: Influence of each adversarial strategy on performance

Results
HSD can be improved by fine-tuning on difficult instances. We have increased robustness while keeping the same level of performance on typical instances from biased Web datasets (see Fig. 5). Mixing support strategies increases their effectiveness.

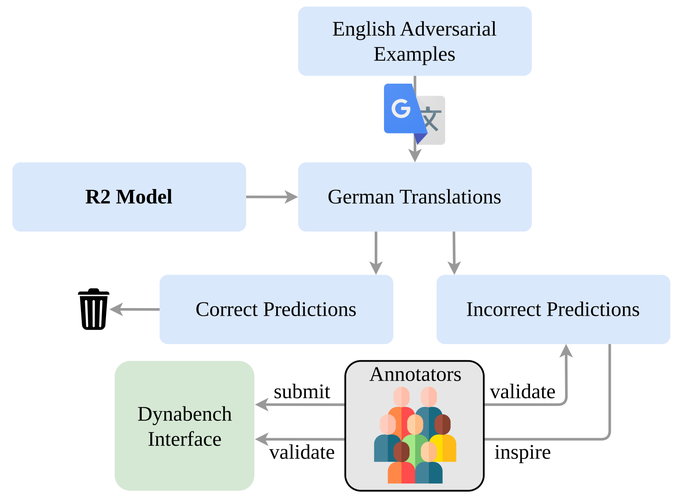

Fig.4: Translating English adversarial examples and manually verifying incorrect predictions. The incorrect predictions and further attacks inspired by them are added to the training data.

3) Mapping associations of religion & spirituality (with P11)

Objective
Bottom-up definitions of religion, spirituality, faith are difficult. We follow Neubert (2016) with a radically discourse-based definition, in the spirit of Wittgenstein, where the meaning of a word is its use. The loud marketplace of Twitter show contests and associations.

Methods
We collected over 100,000 tweets on *religion* and *spirituality.*
We added the pseudowords *sentipositive* and *sentinegative* to each tweet, using automatic sentiment detection. The method clusters words with similar distribution profiles.

Results
An association cloud grouping words with similar distribution is shown in Fig. 1. Positive and negative associations are very strong (see interdisciplinary panel)

People involved in P8
Martin Volk, Gerold Schneider, Janis Goldzycher
Department of Computational Linguistics

Selected Publications

Goldzycher, Janis, Paul Röttger and Gerold Schneider. (2024). Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset. In *Proceedings of NAACL*.

Goldzycher, Janis and Gerold Schneider. (2022). Hypothesis Engineering for Zero-Shot Hate Speech Detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, 75–90, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Schneider, Gerold. (Monograph, 2024, in print). *Text Analytics for Corpus Linguistics and Digital Humanities: Simple R scripts and Tools.* Bloomsbury.

Schneider, Gerold (fc.). Combining Collocation Measures and Distributional Semantics to Detect Idioms. Submitted.