HICC: A Dataset for German Hate Speech in Conversational Context

Lars Schmid[1], Pius von Däniken[1], Patrick Giedemann[1], Don Tuggener[1], Judith Bühler[2], Maria Kamenowski[2], Katja Girschik[2], Laurent Sedano[2], Dirk Baier[2], Mark Cieliebak[1]

[1]ZHAW Centre for Artificial Intelligence,
[2] ZHAW Institute of Delinquency and Crime Prevention
Correspondence: shmr@zhaw.ch

Abstract

We present HICC: A dataset of German **H**ate speech comments **I**n **C**onversational **C**ontext. The dataset comprises the target hate speech comments to be detected, the root post that starts the conversation, and all direct replies to the root that were published before the target. We especially focus on comments that are difficult to identify as hate speech in isolation, but need context for classification. We show that LLM-based classification can successfully incorporate conversational context to identify such comments that do not explicitly contain hateful language. Our dataset is gathered from the social media platform *X* (formerly Twitter), manually annotated, and made freely available for research[1].

Content Warning: This paper contains examples of hate speech for research purposes.[2]

1 Introduction

Research on detecting hateful and toxic speech in social media often focuses on the message level, analyzing messages without their conversational contexts. However, comments on social media are naturally embedded in a conversational context. We show that incorporating this context in message-level analysis of social media comments is beneficial to detecting hateful and toxic speech that does not feature hateful language on the text surface. To this end, we collect and manually annotate a dataset for identifying hateful and toxic comments in German on the platform *X* (formerly Twitter).

Example 1[3] illustrates a target that appears innocuous in isolation, but when read in context, the question functions as a sarcastic dismissal aimed at a political out-group. Humans label it as "toxic speech", and LLMs likewise flip from `0` (no context) to `1` when given the conversational context. The example shows that toxicity can be conveyed through pragmatic cues (sarcasm, insinuation) rather than explicit slurs, and that those cues often reside outside the target span.

This paper makes three contributions. First, we release HICC, a German dataset pairing target comments with their conversational context (Section 3). Second, we propose an evaluation pipeline that first ranks targets by off-the-shelf moderation scores *without* context and then measures how much LLMs recover missed cases *with* context (Section 4). Third, we show recall gains when supplying context – especially for the hardest deciles where context-agnostic tools[4] assign low scores – reaching improvements up to 19 percentage points (pp) (Section 5).

Example 1: Context reveals latent toxicity (id 15878)

Root post: *"This phenomenon was confirmed in a study, here is an excerpt: "A wider road produces ... increased demand – and leads to people using their cars more often than before, creating a new, larger traffic jam."*
Comment 1: *"The Greens and other brainless people actually believe that if there were fewer roads, there would be fewer cars! They just don't realise that the cars simply drive through the villages. But yes, they mostly live in cities with trams etc. where cars are banned."*
Comment 2 (Target): ***"A green study?"***
Moderation endpoints avg. (Target): 0.019
Human label (Target): TOXIC SPEECH
Model verdicts (no context): GPT-4o-mini, Grok-3-mini, Claude 3 Haiku = 0
Model verdicts (+Full): GPT-4o-mini, Grok-3-mini, Claude 3 Haiku = 1
Why it flips. Without any context, the statement *"A green study?"* appears to be an innocuous, neutral question. With context, it becomes clear that the remark is a politically-charged jab and sarcastic dismissal of the study on partisan grounds, especially after a preceding reply calls the Greens 'brainless'.

[1] https://github.com/larscarl/hicc

[2] These examples are necessary to illustrate the challenges in detecting and understanding harmful content. The authors do not condone any of the language or sentiments expressed in them.

[3] See Appendix A.1 for additional examples.

[4] Perspective API and OpenAI's Moderation; see Section 4

2 Related Work

Comprehensive surveys cover abusive language detection broadly (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Jahan and Oussalah, 2023); here we focus on datasets and tasks that incorporate conversational context. Qian et al. (2019) use conversational threads from Reddit and focus on generating counter-speech, but not the identification of hate speech. Vidgen et al. (2021) present a substantially sized dataset comprised of Reddit comment threads annotated with an in-depth taxonomy of abusive language. They include annotations to indicate whether a comment needs context to be interpreted as being abusive. However, they do not consider context in their automatic classification experiments. Demus et al. (2022) present a dataset of toxic speech based on German posts on *X*. They analyze the impact of hateful comments on subsequent posts, but do not include the conversational context in their classification experiments.

Work on whether conversational context improves classification accuracy has not yielded a decisive answer (Pavlopoulos et al., 2020; Xenos et al., 2021; Yu et al., 2022; Pérez et al., 2023). Different approaches use divergent definitions of context (mostly the parent comment), annotation guidelines, label sets, and data sources. One clear exception is the work of Pérez et al. (2023) who collect a dataset of Spanish (Rioplatense variant) tweets and show clear improvements when incorporating the original tweet into classification.

By contrast, the main focus of our work is not whether context helps in general, but whether we are able to correctly classify hateful comments that are not detectable as such in isolation. Specifically, we are interested in online comments that do not explicitly feature offensive words, but can only be interpreted as hateful with the inclusion of the conversational context (Klenner, 2018; Wiegand et al., 2021). Our work is hence closely related to Xenos et al. (2021) who devised a context-sensitivity estimation task. They consider the parent comment of a target comment as the context. Somewhat surprisingly, they found that about one third of the comments perceived as toxic in isolation were rendered neutral when annotators were shown the parent comments. They showed that Perspective API[5] also benefits from including parent posts for comments where the human annotations differ when including the parent posts. Furthermore, they trained a separate system to predict whether a comment requires context to be correctly annotated and showed that this system performs significantly better than randomly sampling comments.

Our work aims to provide a dataset of German hate speech with a focus on integrating the conversational context and to examine its impact on the detection of hateful comments that do not feature obvious hateful language.

3 Data collection

Searching and collecting hateful comments often involves querying with known hateful words and key-phrases (Qian et al., 2019; Yu et al., 2022; Demus et al., 2022), which can bias samples (Schmidt and Wiegand, 2017; Wiegand et al., 2018). HICC was collected organically as part of the project "Social Influencer:in"[6] aiming at evaluating the effectiveness of different counter-speech methods: digital streetworkers browsed news and social media to identify hateful or toxic comments.

A total of 13 digital streetworkers were recruited from undergraduate programs in social work and communication sciences and were deployed and supported over a four-month period.[7] Prior to collection, they completed a six-session training that standardized annotation decisions (definitions and criteria for hate vs. toxic speech, legal framing, search and classification procedures, discourse risks of counter-speech, and onboarding to the technical workflow).

Finding targets. The streetworkers did not search via hate-related keywords or hashtags. Instead, they practiced targeted browsing of sources that routinely elicit polarized comment sections. Concretely, they (i) monitored posts and news framed around groups frequently attacked online (e.g., refugees and migrants, queer people, Muslims, Jews, Sinti and Roma, women, trans and non-binary people, people with disabilities, people affected by poverty, unemployed or homeless people) as well as public figures (politicians across the spectrum, journalists, activists, scientists); and (ii) regularly visited accounts and pages of political parties and politicians (left & right), online

[5] https://www.perspectiveapi.com/ provides a toxicity score.

[6] https://www.zhaw.ch/de/forschung/projekt/74272

[7] Their ages ranged from 21 to 36. Their activities lasted from September to December 2024. Ten of them received a monthly salary; three received ECTS credits. The weekly workload was around 6 hours.

newspapers, journalists, activist collectives, anti-discrimination actors, religious organizations, and topic influencers (e.g., body positivity, climate, anti-racism, queer). They followed or subscribed to these sources; over time, platform recommendation systems increasingly surfaced similarly polarized threads. This strategy avoids lexical keyword/hashtag bias while focusing on realistic venues for counter-speech.

Use of context. Annotators read the surrounding conversation before deciding whether a post warranted intervention, inspecting the root post and the ongoing discussion (main thread and subthreads) to assess stance, sarcasm, and implicit attacks, and then usually selecting one target comment per thread (either the root or a reply)[8]. For release and experiments, we reconstruct context as the root post and all direct replies to the root that were published before the target. Deeper branches are excluded.

Annotation tooling. A custom browser extension highlighted texts in the current tab according to their likelihood of toxicity or hate and served as the data-collection interface[9]. Annotators labeled the chosen target as one of: (i) "needs intervention (hate)"[10], (ii) "needs intervention (toxic)"[11], (iii) "intervention not recommended", (iv) "no intervention needed" [12]. The extension stored the annotation, URL, and page HTML at capture time. For evaluation, we treat (i) as *hate speech* and (ii) as *toxic speech*; (iii)-(iv) are excluded unless stated otherwise.

The integrated classifier was a SwissBERT model (Vamvas et al., 2023) with the *de_ch* adapter (Pfeiffer et al., 2022), finetuned on DETOX (Demus et al., 2022). It predicted five independent[13] binary targets (*toxic*, *hate speech*, *legal relevance*, *threat*, *extremism*) and shaded comments according to the maximum probability over these labels. On DETOX, the classifier shows moderate precision on hate and conservative recall on toxicity (see Appendix A.3 for details).

Our collection protocol yields *positive* supervision ("needs intervention (hate/toxic)") but not exhaustive negatives. Streetworkers were not instructed to label every comment in a thread – neither all non-problematic nor all problematic ones. The few "no intervention needed" flags mostly reflect occasional false positives of the classifier and are too sparse/inconsistent as true negatives. Likewise, they did not always annotate every hateful comment; often several hateful replies remained unlabeled because one was chosen arbitrarily for intervention. "Intervention not recommended" is not a negative toxicity/hate label. Consequently, the dataset does not support training or evaluating a binary "requires intervention vs. no intervention" system without additional annotation.

In total, the HICC dataset contains 21'338 tweets: 1'812 annotated targets (940 labeled as "needs intervention (toxic)", 734 "needs intervention (hate)", 108 as "intervention not recommended", and 30 "no intervention"[14]). We combine each target with its root post and preceding direct replies to the root. The resulting contexts have a mean size of 11.8 posts per target. On average, the target comments have 10 previous comments (standard deviation of 7.65) under the root post.

To comply with *X*'s Developer Policy[15] we release Post IDs for all nodes plus annotations and derived metadata. Content can be reconstructed via the official API subject to access rights.

4 Experiments

The purpose of our experiments is to identify comments that are not easily detectable as hate speech without context and then evaluate whether LLM-based classification is able to identify them when provided with appropriate context. We do so by ranking and binning comments that were manually

[8] The choice of target was often ad hoc. Streetworkers did not systematically pick the "worst" comment, and a thread could contain several unlabeled but clearly hateful replies.

[9] The on-page classifier operated *per comment* on isolated text only; no hashtag, page-, or thread-level signals. Annotators could select unhighlighted items.

[10] Hate speech is defined as explicit or implicit communicative acts against groups or individuals along (presumed) group membership that devalue, discriminate, or insult; includes slurs, dehumanization, identity-based attacks, or incitement to hatred/violence (see Appendix A.2).

[11] Toxic Speech is defined as language that degrades the conversational climate by attacking, belittling, harassing, exerting pressure, or being destructive/manipulative; measured by its potential to provoke aggressive responses or drive others away (see Appendix A.2).

[12] see Appendix A.2 for the exact definitions used as well as German originals

[13] One-vs-rest heads; labels may co-occur.

[14] This low number is due to the streetworkers not being encouraged to explicitly find non-problematic comments but occasionally flagging false positives of the classifier.

[15] X Developer Policy: "Content redistribution" – you may only distribute Post IDs, Direct Message IDs, and/or User IDs. Academic researchers may share unlimited Post/User IDs for non-commercial research. See: https://developer.x.com/en/support/x-api/policy.

annotated as being hateful. The ranking is based on hatefulness and toxicity scores from existing content moderation tools described in Section 4.1. Binning the ranking allows us to identify sets of comments that are scored as having a low probability of being hateful by the existing moderation tools. We then measure the impact of incorporating context specifically for those low-ranked bins when applying LLM-based classification.

Our experimental pipeline thus has two stages: (i) automatic moderation scoring for candidate ranking and binning, and (ii) LLM evaluations that vary conversational context.

4.1 Automatic moderation scores.

Comments in HICC that have been labeled as either "needs intervention (hate)" or "needs intervention (toxic)" are scored by three off-the-shelf moderation systems: (1) *Perspective API* (`toxicity`, `severe_toxicity`); (2) the *OpenAI* moderation endpoint[16] (`hate`); and (3) our own classifier (see Section 3) (`toxicity`, `hate_speech`). Crucially, this step does not involve any conversational context, i.e., only the target comment is submitted. We retain each raw score returned by the moderation systems and compute an *aggregate average* per comment (denoted `aggregate_avg`). The scores are never treated as ground-truth labels; rather, we use them to sort the comments to identify comments that are difficult to classify using context-agnostic methods (i.e. comments manually annotated as hate speech that have a low aggregate score based on the moderation endpoints).

4.2 LLM-based classification with context variations.

We apply three current models, GPT-4o-mini[17], Grok-3-mini[18], and Claude 3 Haiku[19] as classifiers using the same prompts. For each comment we ask two questions, *"Is the following comment toxic speech?"* and *"Is the following comment hate speech?"*, preceded by definitions of the respective concept.[20] The query is repeated under three context conditions[21]:

(1) **No Context ("No")** – the target text in isolation.

(2) **Root Post ("+Orig")** – the root post and the target.

(3) **Full Prior Context ("+Full")** – the root post and all direct replies to the root published before the target, followed by the target[22].

Models return a free-form rationale plus a binary verdict (`0`=No, `1`=Yes). With 3 models $\times$ 2 prompts $\times$ 3 contexts, we collect 18 verdicts per comment.

5 Results

To evaluate how well the LLM-based classification is able to identify hateful comments of various levels of explicitness, we compare model verdicts against the streetworkers' *target* annotations. By design, annotators typically selected one intervention target per conversation thread – after reading the surrounding discussion – so the dataset contains positive labels for targets only. In some cases multiple streetworkers intervened in the same thread, resulting in more than one labelled target. All other posts in the same thread are *unlabeled* (they are not confirmed negatives), even if they may also satisfy hate/toxicity criteria. Because negatives are not exhaustively annotated, we report *Recall* on the target set: a true positive (TP) when the model flags the target under the matching prompt (hate vs. toxic), and a false negative (FN) otherwise. We do not estimate Precision or Accuracy on the full thread, as true negatives and false positives cannot be derived from the unlabeled context.[24]

Main Experiments Table 1 reports Recall for the three LLMs under the three context conditions described in Section 4.2. Context improves recall consistently, with the largest gains on the combined *Overall Recall* (combining the results of Hate and Toxic comments[25]) when the full conversational thread is provided. Across all models, the average *Overall Recall* rises from 88-89% (no context) to 90-93% (full prior context). GPT-4o-mini gains the most (+3.4 pp), followed by Grok-3-mini (+3.3 pp) and Claude 3 Haiku (+1.2 pp).

[16] https://platform.openai.com/docs/guides/moderation

[17] https://platform.openai.com/docs/models/gpt-4o-mini

[18] https://docs.x.ai/docs/models

[19] https://docs.anthropic.com/en/docs/about-claude/models/all-models

[20] Full definitions are provided in Appendix A.2.

[21] For brevity, we use the shorthands **No**, **+Orig**, and **+Full** in all tables and figures.

[22] Replies posted after the target are excluded.

[23] `aggregate_avg` = mean of moderation scores from Perspective (`toxicity`, `severe_toxicity`), OpenAI Moderation (`hate`), and our SwissBERT classifier (`toxicity`, `hate_speech`), computed on the *target comment only* (no context). Lower values indicate comments that are harder for context-agnostic tools.

[24] We provide a Precision analysis of our LLM-based classifier on a different corpus in Appendix A.3.

[25] Computed by $\frac{TP\_tox+TP\_hate}{TP\_tox+TP\_hate+FN\_tox+FN\_hate}$

Model	Hate Recall (%)			Toxicity Recall (%)			Overall Recall (%)		
	No	+Orig	+Full	No	+Orig	+Full	No	+Orig	+Full
GPT-4o-mini	89.9	89.1	**91.5** (+1.6)	88.7	92.0	**93.4** (+4.7)	89.2	90.8	**92.6** (+3.4)
Grok-3-mini	**89.9**	88.8	89.3 (−0.6)	86.5	92.6	**92.6** (+6.1)	87.9	91.0	**91.2** (+3.3)
Claude 3 Haiku	94.7	93.9	**95.2** (+0.5)	85.2	86.5	**86.9** (+1.7)	89.2	89.6	**90.4** (+1.2)

Table 1: Recall on the HICC dataset. "+Orig" = root post provided; "+Full" = full prior context (root post plus all direct replies before the target).

Difficulty Bin (`aggregate_avg` range[23])	n	GPT-4o-mini		Grok-3-mini		Claude 3 Haiku	
		Δ +Orig	Δ +Full	Δ +Orig	Δ +Full	Δ +Orig	Δ +Full
(0.0145, 0.0915]	102	+8.1	**+15.2**	**+19.2**	**+19.2**	+10.1	**+16.2**
(0.0915, 0.168]	94	+8.9	**+13.3**	**+10.0**	+8.9	−3.3	−2.2
(0.168, 0.244]	132	+1.6	**+3.2**	+2.4	**+3.2**	0	−0.8
(0.244, 0.32]	124	−0.9	0	**+1.8**	+0.9	**+1.8**	**+1.8**
(0.32, 0.396]	172	0	0	−3.2	−1.3	−2.5	−2.5
(0.396, 0.472]	136	0	0	0	0	0	0
(0.472, 0.548]	125	−1.9	0	−0.9	−0.9	0	0
(0.548, 0.625]	60	−1.9	−1.9	0	0	−1.9	0
(0.625, 0.701]	44	0	0	0	0	0	0
(0.701, 0.777]	16	0	0	0	0	0	0

Table 2: **Per-decile recall gains from adding conversational context.** Cells report Δ recall in percentage points relative to the *No context* baseline; **bold** marks the larger gain between +Orig and +Full within a row. Lower bins correspond to harder cases (lower moderation scores). Context helps most in the two hardest bins (up to +19 pp), and has negligible or slightly negative effect once toxicity is already evident.

A complementary cross-label analysis (Appendix A.4) shows a strong asymmetry: a single toxicity prompt nearly recovers all hate cases ($\geq 97.6\%$ recall), but the reverse is much weaker (70-82%).

Recall by difficulty bin. Table 2 drills down on *Overall Recall* along the `aggregate_avg` spectrum that we use as a proxy for "moderation difficulty". That is, we sort the comments according to the aggregate average and then assign them to 10 bins. Context helps most in the hardest bins (`aggregate_avg` ≤ 0.168), boosting recall by up to 19 pp, while having negligible or occasionally negative effect once toxicity is already salient according to the moderation endpoints.

6 Conclusion

We have presented HICC, a dataset for German hate speech detection curated to emphasize the importance of conversational context. We showed that providing the full prior context (root post + prior direct replies) to LLMs leads to improved Recall in identifying both toxic and hate speech, with overall recall increasing from approximately 88-89% without context to 90-93% with full prior context. The most substantial gains in Recall, up to 19 percentage points, were observed for comments that were initially ranked as having a low probability of being hateful by context-agnostic moderation tools, highlighting the value of context in detecting subtle or implicit hate speech. Future work will expand the dataset and include true negatives to be able to evaluate precision. Our recall-based evaluation indicates that the approach is beneficial when assisting streetworkers in identifying hateful comments that require context. The dataset is freely available for research purposes[26].

Funding and project partners

This work was supported by *Nationaler Aktionsplan zur Verhinderung und Bekämpfung von Radikalisierung und gewalttätigem Extremismus*, *Digitalization Initiative of the Zurich Higher Education Institutions (DIZH)*, *Lotteriefonds Appenzell Ausserrhoden*, *Swisslos-Fonds Kanton Solothurn*, *Eidgenössische Kommission gegen Rassismus (EKR)*, *NCBI Schweiz*, and *ZHAW School of Management and Law*.

References

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022.

[26] https://github.com/larscarl/hicc

DeTox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Manfred Klenner. 2018. Offensive language without offensive words (OLWOW). In *14th Conference on Natural Language Processing KONVENS 2018*, page 11.

John Pavlopoulos, Jeffrey Scott Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *ArXiv*, abs/2006.00998.

Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and 1 others. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding-Royer, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Conference on Empirical Methods in Natural Language Processing*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Z. Margetts, Patrícia G. C. Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *North American Chapter of the Association for Computational Linguistics*.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*, page 1.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Examples

Examples 2 and 3 illustrate *how* and *why* conversational context can overturn toxicity or hate speech assessments. Each boxed example pairs the comment with its surrounding thread, reports both human annotations and model verdicts under different context windows, and ends with a short rationale for the observed flip.

Example 2: Context flips Schadenfreude into detectable toxicity (id 7436)

Content Warning: Sexual Assault
Root post: *"Woman (32) forced to perform sexual acts in flat in Zug: Alleged perpetrator (26) found dead after falling through a window."* [a]
Comment: *"At least he left afterwards"* [b]
aggregate_avg: 0.019
Human label: TOXIC SPEECH
Model verdicts (no context): GPT-4o-mini, Grok-3-mini, Claude 3 Haiku = 0
Model verdicts (+Full): GPT-4o-mini, Grok-3-mini, Claude 3 Haiku = 1

Why it flips. The short German phrase contains no obvious slur or profanity. Only when the headline reveals that it refers to a suspected rapist who died does the comment read as a celebration of lethal harm (*Schadenfreude*). With that context, both humans and LLMs judge it as toxic speech.

[a] Original statement in German: "Frau (32) in Zuger Wohnung zu sexuellen Handlungen gezwungen: Mut-

masslicher Täter (26) nach Fenstersturz tot aufgefunden."

[b]Original statement in German: "Wenigstens ist er danach gegangen"

Example 3: Scapegoating migrants through implicit blame (id 15857)

Root post: *"[ANONYMIZED]: The case of a fourteen-year-old contract killer in Marseille shakes France..."* [a]
Comment: *"That's just the way it is with unchecked immigration."* [b]
aggregate_avg: 0.127
Human label: HATE SPEECH
Model verdicts (no context): GPT-4o-mini = 0, Grok-3-mini = 0, Claude 3 Haiku = 1
Model verdicts (+Orig): GPT-4o-mini = 1, Grok-3-mini = 1, Claude 3 Haiku = 1
Model verdicts (+Full): GPT-4o-mini = 1, Grok-3-mini = 0, Claude 3 Haiku = 1
Why it flips. In isolation the sentence passes as policy critique. Once paired with the tweet about a contract killing, it implicitly blames migrants for violent crime, crossing the hate-speech boundary for most models.

[a]Original statement in German: "[ANONYMIZED] on *X*: Der Fall eines vierzehnjährigen Auftragskillers in Marseille erschüttert Frankreich…"

[b]Original statement in German: "so ist es halt mit ungebremster Zuwanderung"

A.2 LLM prompts

Here we list our working definitions of hate speech and toxic speech and give an example prompt used in our LLM-based experiments.

A.2.1 Definition 'Hate Speech'

Below, we provide our working definitions of hate speech; first in German, then translated to English.

Hate Speech Definition (German)

Unter Hate Speech werden kommunikative Handlungen oder Verhaltensweisen gegen Gruppen oder Einzelpersonen mit dem Ziel der (bewussten oder unbewussten) Abwertung, Diskriminierung oder Beleidigung derselben verstanden. Dabei kann die Abwertung aufgrund verschiedener Merkmale (bspw. Religion, Ethnizität, Nationalität, Hautfarbe, Abstammung, Geschlechtsidentität, sexuelle Orientierung, Alter, Behinderung, Körperform, Krankheit, ökonomische Situation, Beruf, Bildungshintergrund, politische Orientierung und weiteres) offensichtlich oder auch verdeckt sein. Hate Speech umfasst Beleidigungen, Verleumdungen ebenso wie Aussagen, die abwertende Stereotype fördern und zu Hass und Gewalt gegen Personen entlang von (vermuteten) Gruppenzugehörigkeiten aufrufen. Hate Speech kann sich direkt oder indirekt äussern.

Erweiterte Hate Speech Definition, die im Projekt verwendet wird:

- Verwendung von Schimpfwörtern
- Abwertung der Identität / Angriffe auf die Identität
- Aufrufe zu Gewalt und Drohung
- Entmenschlichende Sprache
- Verbreitung von Falschinformationen
- Sexuelle Belästigung
- Mobbing
- Kann strukturelle Diskriminierung enthalten (Geschlecht, Hautfarbe, Religion, Herkunft/Ethnizität, Sprache, Sesshaftigkeit, Gesundheit/Körper, Besitz/Klasse, Bezug globaler Norden und globaler Süden, Art der Gesellschaft (modern/fortschrittlich))
- Kann Freund-Feind-Denken beinhalten (z. B. Sport, Vereine, Fans, Aktivitäten (Velo-Auto/Klima, Wähler:innen einer Partei, etc.))
- Kann sich gegen Berufsgruppen richten (z. B. Medienschaffende, Politiker:innen, Behörden, Polizei, Vertreter:innen einer Organisation, etc.)
- Individuell, persönlicher Konflikt (Beleidigung, Beschimpfung, üble Nachrede, Nötigung, Drohung)

Keine Hassrede ist:

- Kritik – auch wenn sie harsch ist
- Freie Meinungsäusserung
- Nicht wertende Verallgemeinerung
- Fremdsprache
- Dialekt
- Falschinformation

Hate Speech Definition (English translation)

Hate speech refers to communicative actions or behavior against groups or individuals with the aim of (consciously or unconsciously) devaluing, discriminating or insulting them. The devaluation based on various characteristics (e.g. religion, ethnicity, nationality, skin color, ancestry, gender identity, sexual orientation, age, disability, body shape, illness, economic situation, profession, educational background, political orientation and others) can be obvious or hidden. Hate speech includes insults and slander as well as statements that promote derogatory stereotypes and incite hatred and violence against people based on (presumed) group affiliations. Hate speech can be expressed directly or indirectly.

Extended hate speech definition used in the project:

- Use of swear words
- Devaluation of identity / attacks on identity
- Incitement to violence and threats
- Dehumanizing language
- Dissemination of false information
- Sexual harassment
- Bullying
- May include structural discrimination (gender, skin color, religion, origin/ethnicity, language, sedentariness, health/body, property/class, reference global north and global south, type of society (modern/progressive))
- Can include friend-foe thinking (e. g. sports, clubs, fans, activities (bike-car/climate, voters of a party, etc.))
- Can be directed against professional groups (e.g. media professionals, politicians, authorities, police, representatives of an organisation, etc.)
- Individual, personal conflict (insult, abuse, defamation, coercion, threat)

No hate speech is:

- Criticism - even if it is harsh
- Free expression of opinion

- Non-judgemental generalisation
- Foreign language
- Dialect
- Misinformation

A.2.2 Definition 'Toxic Speech'

Below, we provide our working definitions of toxic speech; first in German, then translated to English.

Toxic Speech Definition (German)

Toxic Speech bezeichnet eine Sprache, die das Gesprächsklima vergiftet und den Diskurs für Hass und sprachliche Gewalt öffnet. Toxic Speech meint daher Aussagen, die Gesprächspartner:innen vor den Kopf stossen, sie blossstellen, die angriffig sind, die Druck ausüben oder destruktiv und manipulativ sind. Die Toxizität gibt an, wie "giftig" ein Kommentar für eine Konversation ist. Regt er zu negativen/aggressiven Reaktionen an oder würde man ihn vielleicht einfach unkommentiert lassen? Je stärker er zu aggressiven Äusserungen oder gar zum Austritt anderer Teilnehmer aus der Konversation anregt, desto toxischer ist der Kommentar.

Toxic Speech Definition (English translation)

Toxic speech refers to language that poisons the climate of dialogue and opens up the discourse to hatred and linguistic violence. Toxic speech therefore refers to statements that offend, expose, attack, exert pressure or are destructive and manipulative. Toxicity indicates how 'poisonous' a comment is for a conversation. Does it provoke negative/aggressive reactions or would you simply leave it uncommented? The more it encourages aggressive comments or even the withdrawal of other participants from the conversation, the more toxic the comment is.

A.2.3 Full Prompt (example)

Here, we show an example prompt for LLM-based classification including the full conversational context in German.

Full Prompt Example (German)

`System:` Sie sind ein Experte für die Moderation von Inhalten mit umfangreicher Erfahrung in der Erkennung von Hate Speech. Ihre Aufgabe besteht darin, von Nutzern bereitgestellte Kommentare auf der Grundlage der folgenden Definition von Hate Speech zu bewerten:

[Hate-Speech-Definition]

Wenn Sie einen Kommentar analysieren, müssen Sie ein binäres Urteil - 0 für 'ist nicht Hate Speech' und 1 für 'ist Hate Speech' - zusammen mit einer knappen, klaren Erklärung Ihrer Argumentation abgeben. Achten Sie darauf, dass Sie sich streng an das strukturierte Format halten, das im HateSpeechResponse-Schema definiert ist. Seien Sie objektiv, verweisen Sie auf die angegebene Definition, wenn dies angebracht ist, und vermeiden Sie unnötige Kommentare.
Wichtig: Wenn zusätzlicher Kontext wie ein ursprünglicher Kommentar oder ein Gesprächsverlauf bereitgestellt wird, berücksichtigen Sie diesen bei Ihrer Bewertung. Ein Kommentar kann in einem Kontext als Hate Speech gelten, während er in einem anderen Kontext harmlos sein könnte. Beziehen Sie den Kontext explizit in Ihre Begründung ein.
`Human:` Ist folgender Kommentar Hate Speech?

> 'die vollständige woke Verblödung sehen wir hier in Wort und Bild, von einer oder als was sie sich auch immer bezeichnet, die in BS Genderstudies absolviert hat!'

Der Kommentar ist eine Antwort auf diesen ursprünglichen Kommentar:

> '[ANONYMIZED] auf *X*: „«Es gibt zwei Geschlechter. #Mann und #Frau.» - diesen Satz wollen die woken Linken strafbar machen und in die «Anti-#Rassismus-Strafnorm» integrieren. Komplett irr. Ein Frontalangriff auf die Natur und unsere #Meinungsfreiheit. #only2'

Hier ist der vorherige Gesprächsverlauf:

- Kommentar 1: 'Du solltest da eigentlich etwas tolerenter sein. Wer [ANONYMIZED] als zweites Ego hat, sollte sich nicht allzu weit aus dem Fenster lehnen.'
- Kommentar 2: 'Dir hat die Natur wirklich übel zugesetzt.'

System / Label	Precision	Recall	F1
DETOX GBert – Hate	0.78	0.67	0.71
DETOX XLM-R – Toxicity	0.56	0.56	0.54
DETOX Avg. – Hate	0.66	0.65	0.62
DETOX Avg. – Toxicity	0.40	0.40	0.39
Ours – Hate	0.51	0.61	0.56
Ours – Toxicity	0.56	0.16	0.25
GPT-4o-mini – Hate	0.32	0.79	0.46
GPT-4o-mini – Toxicity	0.08	1.00	0.15

Table 3: Evaluation on DETOX dataset. DETOX GBert and XLM-R are the best results reported in Demus et al. (2022).

A.3 Precision Evaluation on DETOX

To gauge precision of the LLM-based classification, we evaluated GPT-4o-mini on the German DETOX corpus (Demus et al., 2022) which does not provide conversational context. The definitions for hate speech and toxic speech used in the prompt were taken from Demus et al. (2022) to ensure comparability. Table 3 compares the LLM against our SwissBERT baseline (threshold 0.5).

While GPT-4o-mini recalls all toxic instances, its precision is considerably lower than our SwissBERT model's, highlighting a recall-precision trade-off that warrants future mitigation.

Additionally, we performed a preliminary Precision evaluation using the 30 available instances of HICC that were labeled as "no intervention" and measured a Precision of 60% for the hate speech detection prompt and 20% for the toxic speech detection prompt using GPT-4o-mini when providing the full context. This result provides a first indication of the performance of the LLM-based classification, but the sample size is too small to

draw conclusions. Discussions with the streetworkers revealed that they used the label inconsistently and confounded it with the "intervention not recommended" label.

A.4 Cross-Label Experiments

We tested how robust each model is when the prompt/label pair is mismatched. Table 4 shows recall for the *Toxic-on-Hate* and *Hate-on-Toxic* settings using the full-context condition.

Model	Tox→Hate	Hate→Tox
GPT-4o-mini	99.5	74.0
Grok-3-mini	98.9	70.7
Claude 3 Haiku	97.6	81.8

Table 4: Cross-label recall (%) in the full-context condition.

Across all three models, a single toxicity prompt almost fully covers hate-speech instances ($\geq$ 97.6 %), whereas the inverse direction is much weaker (70-82%). This suggests that toxic language is a superset: a single toxicity prompt nearly covers all hate cases, but not vice-versa.