nymera@GermEval Shared Task 2025: One Ensemble, Many Harms: A Unified Transformer Approach to Harmful Content Detection in German Social Media

Hanna Köpcke
University of Applied Sciences Mittweida
Mittweida, Germany
koepcke@hs-mittweida.de

Abstract

We present a unified approach to detecting harmful content in German social media, developed for the GermEval 2025 Shared Task: Harmful Content Detection in Social Media (Felser et al., 2025). Our system addresses all three subtasks—(1) calls to action inciting harmful acts (C2A), (2) violence glorification (VIO), and (3) attacks on the free democratic basic order (DBO)—using a single ensemble-based framework. We fine-tuned a three-model ensemble (GBERT-large, XLM-RoBERTa-base, and DeBERTa) for each subtask and aggregated their predictions through soft-voting. To mitigate severe class imbalance in the training data, we augmented the dataset with synthetic examples and manually relabeled instances for minority classes, and applied oversampling during training. This unified modeling approach achieved strong performance in the official evaluation: our system obtained macro-F1 scores up to 0.83, ranking 1st in the DBO subtask, 3rd in C2A, and 4th in VIO.

1 Introduction

Harmful online content has evolved beyond generic hate speech to include nuanced categories such as explicit calls to action (incitements), violence glorification, and attacks on democratic institutions. These forms of content pose serious risks: direct calls to violence can spur real-world harm, and anti-democratic narratives erode public trust in societal structures. In response, the GermEval 2025 Shared Task on Harmful Content Detection in Social Media (Felser et al., 2025) introduced three focused subtasks for German: detecting calls to action inciting harmful acts (C2A), violence glorification (VIO), and attacks on the free democratic basic order (DBO). All task datasets consist of German Twitter posts from an extremist network collected in 2014–2016, providing a realistic testbed for these phenomena. Each subtask targets a distinct semantic layer of harmful discourse: C2A detection hinges on recognizing illocutionary intent (when users urge others to act), VIO involves identifying rhetoric that praises or justifies violent acts, and DBO requires spotting ideological content that undermines fundamental democratic values. These phenomena are often subtle; for example, a post may implicitly praise violence via humor or historical reference, making them challenging to identify automatically.

A key difficulty across all three categories is extreme class imbalance. In these datasets, harmful posts are vastly outnumbered by benign content (only ∼7–10% of instances are positive in C2A and VIO, and under 1% in the most severe DBO category). This imbalance can lead classifiers to overlook the rare but critical positive instances. Another challenge is pragmatic ambiguity: whether a message is a dangerous incitement or merely hyperbolic sarcasm often depends on subtle context and connotation. Effective solutions thus need to capture context, intent, and domain-specific cues without being overwhelmed by the majority class.

In this paper, we propose a unified ensemble-based method to tackle all three subtasks. Our approach fine-tunes three complementary Transformer models (GBERT-large, XLM-RoBERTa-base, and DeBERTa) for each task and combines their outputs through soft-voting (probability averaging) to produce robust predictions. By leveraging a monolingual German model alongside a multilingual model and an advanced Transformer architecture, the ensemble captures both language-specific nuances and generalizable patterns. We also address the class imbalance via targeted data augmentation and resampling: generating synthetic minority-class examples and adding a handful of manually relabeled misclassified posts, then oversampling the minority classes during training. Using this strategy, our unified system achieved strong

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 357–365, 2025

performance across all three GermEval 2025 subtasks. To support reproducibility, we have made the full implementation publicly available in a GitHub repository.[1]

After presenting the current state of the literature on harmful content detection in Section 2, Section 3 introduces the dataset used in this work. Section 4 then details our unified transformer ensemble methodology and the data augmentation strategies employed. Section 5 presents and discusses the experimental results, and Section 6 concludes the paper with a summary of our findings and some directions for future work.

2 Related Work

Ensemble Methods for Harmful Content Detection: Ensemble classifiers have become standard in hate speech detection, as combining multiple models helps capture diverse abusive language cues and improves robustness. Transformer-based ensembles in particular have achieved state-of-the-art results in hate/offensive content tasks (e.g., winning entries in HASOC and SemEval challenges (Glazkova et al., 2021; Wiedemann et al., 2020)). Recent studies show that ensembling models emphasizing different facets of toxic content can yield further gains over single models (Kucukkaya and Toraman, 2025). Hybrid architectures mixing transformers with other classifiers likewise enhance performance by leveraging complementary strengths (Mazari et al., 2024). These findings establish ensembling as an effective strategy for complex content moderation tasks.

Handling Class Imbalance: Harmful content datasets are typically highly imbalanced, with genuine toxic instances vastly outnumbered by benign content. Without mitigation, classifiers tend to predict the majority class, missing rare but critical cases. Data-level solutions such as oversampling minority classes or generating synthetic examples have been shown to improve the learning of under-represented classes (Yuan and Rizoiu, 2025; Achmann-Denkler et al., 2024). Algorithm-level techniques like class weighting or focal loss (Zhang et al., 2024) also help by biasing the model toward rare categories. In practice, top systems often combine these approaches: for example, the winning GermEval 2021 toxicity system applied aggressive upsampling and cost-sensitive training to boost minority-class recall (Risch et al., 2021). Guided by these insights, we employ oversampling and weighted loss to counter class imbalance in our models.

Multilingual and Multi-Task Modeling: Cross-lingual approaches leverage multilingual transformers (e.g., mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020)) to transfer knowledge from high- resource language corpora to lower-resource languages like German, improving toxic content classification (Muller et al., 2021; Calizzano et al., 2021). At the same time, language-specific models often better capture cultural nuances in abusive content (Zeinert et al., 2021). Multi- task learning across related content moderation tasks can also boost performance by sharing representations of "harmfulness" between tasks (Morgan et al., 2021; Ghosh et al., 2023). In this work, we opt for a simpler unified framework: we fine-tune separate models for each subtask and ensemble their predictions, avoiding the complexity of a multi-task model.

Datasets and Evaluation: The research community has developed numerous datasets and benchmarks for harmful content, though resources for fine-grained categories such as DBO or violence glorification are still limited. For German, the GermEval 2018 shared task (Wiegand et al., 2019) released a corpus of approximately 5,000 tweets annotated for offensive language in both coarse- and fine-grained categories, establishing the first German benchmark for hate speech and profanity detection. The GermEval 2019 shared task (Struß et al., 2019) extended this resource substantially by providing a corpus of about 18,000 similarly annotated tweets. The HASOC (Hate Speech and Offensive Content) track further provided German data in 2019–2022, covering hate, offensive, and profane content in a multilingual setting (Modha et al., 2019; Mandl et al., 2020, 2021; Modha et al., 2022). More recently, specialized datasets have emerged for specific harmful facets: (Krenn et al., 2024) compiled GERMS-AT, ∼8k German news comments annotated for sexism, to enable research on misogyny detection in German. The VioLence corpus (Saha et al., 2023) targets violence-inciting texts. The GermEval 2025 shared task fills an important gap by providing focused datasets for incitement (C2A), violence glorification (VIO), and anti-democratic content (DBO) in German.

[1] https://github.com/hannakoepcke/germeval2025

Subtask	Train	Trial	Test
C2A	6,840	1,051	2,982
DBO	7,454	1,053	3,194
VIO	7,783	1,051	3,335

Table 1: Number of instances in the GermEval 2025 dataset per subtask. Test labels were not provided during the competition.

3 Dataset

For a detailed description of the tasks and datasets, see (Felser et al., 2025). The shared task provided three datasets corresponding to the subtasks C2A, DBO, and VIO, each split into training, trial (development), and test sets, as shown in Table 1. The training sets contain roughly 7–8 k posts each, the trial sets around 1 k posts, and the test sets about 3 k posts with labels withheld for evaluation. All data consists of German tweets from a far-right extremist network collected between 2014 and 2016, written in informal style with frequent slang, spelling variations, emojis, and social-media markers such as hashtags, @mentions, and URLs. C2A and VIO are binary classification tasks, with positive instances being very scarce – only about 9–10% of training posts in C2A and around 7% in VIO are labeled `TRUE`. DBO is a four-class task with the labels *nothing*, *criticism*, *agitation*, and *subversive* and is extremely skewed: the majority *nothing* class comprises roughly 84% of training examples, while the rare *subversive* class accounts for less than 1%. Figure 1 illustrates the class distribution in the training set for each subtask, highlighting these imbalances. The trial sets were constructed as stratified samples of the training data to give participants insight into the data structure, so their class proportions mirror those of the training set and are therefore not shown separately.

4 Methodology

Our approach consists of fine-tuning transformer-based classifiers for each subtask and then combining their predictions through a soft-voting ensemble. We address the severe class imbalance in all three subtasks by a combination of oversampling and data augmentation. Below, we detail our training pipeline and the iterative refinement steps we took to improve performance:

4.1 Baseline Model Training

For each task (C2A, DBO, VIO), we fine-tuned pre-trained transformer models on the task's training data. Each model contributes unique advantages in terms of language coverage or architecture:

- **GBERT-large (Chan et al., 2020):** a BERT-large model trained exclusively on extensive German corpora. As a monolingual German language model, GBERT-large captures German-specific linguistic nuances, leading to state-of-the-art performance on German-language NLP tasks (e.g., document classification and NER).
- **XLM-RoBERTa-base (Conneau et al., 2020):** a multilingual Transformer model based on RoBERTa, pre-trained on text from 100 languages. XLM-RoBERTa offers a broad linguistic scope, enabling effective cross-lingual representations. Notably, it outperforms the original multilingual BERT on various cross-lingual benchmarks, with especially strong gains for low-resource languages. This wide language coverage makes it well-suited for tasks involving diverse or non-English text.
- **DeBERTa (He et al., 2021):** an advanced Transformer architecture (Decoding-enhanced BERT with Disentangled Attention) that introduces novel improvements over BERT/RoBERTa. DeBERTa employs a disentangled attention mechanism, where each token is represented by separate content and position embeddings, and an enhanced mask decoder that incorporates positional information more effectively during pre-training. These architectural innovations boost model performance and efficiency, enabling DeBERTa to achieve state-of-the-art results on numerous NLP benchmarks.

Each model in the ensemble brings complementary strengths: GBERT-large provides deep coverage of German linguistic phenomena, XLM-RoBERTa-base contributes robust cross-lingual understanding, and DeBERTa offers cutting-edge architectural enhancements. By combining these models, our ensemble is able to leverage both language-specific knowledge and generalizable representations, improving overall predictive performance.

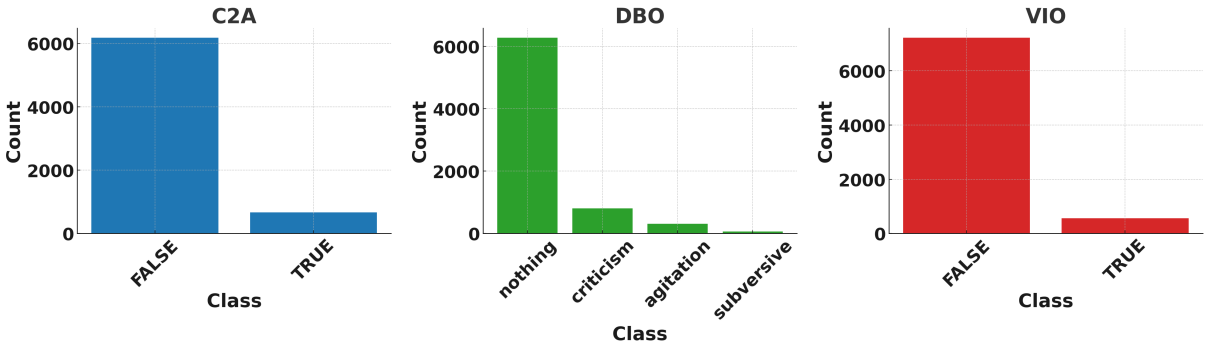

Figure 1: Class distributions across training splits for each subtask

Each model was trained with binary or multi-class cross-entropy as appropriate. We used early stopping on the validation (trial) set to prevent overfitting. The outputs of these individual models were later combined in an ensemble by averaging their predicted probabilities (soft voting), leveraging the complementary strengths of different model architectures. This ensemble strategy is lightweight yet robust, and ensemble methods have proven effective in prior hate speech tasks (Kucukkaya and Toraman, 2025).

4.2 Handling Class Imbalance

The training data for all subtasks was highly skewed (e.g. only 7–10% positives in C2A and VIO, and <1% for the subversive class in DBO). To ensure the models learned minority classes, we applied random oversampling of the minority class instances. Additionally, we employed a class-weighted loss function to further bias training toward the minority classes. In practice, we duplicated examples from the under-represented classes until approximately balancing the class distribution. Importantly, any additional data (described next) was added before oversampling was performed, so that the oversampling procedure could incorporate these new examples as well. By oversampling after augmentation, we ensured the final effective training set had a more balanced mix of classes, mitigating the bias toward predicting the majority class.

4.3 Iterative Data Augmentation

After the initial training run, we analyzed the model's predictions on the test set to identify patterns of errors or obvious gaps in the training data. We found that the models struggled with some subtle positive instances due to the paucity of examples. To address this, we augmented the training set with additional examples for each subtask, using two strategies:

Synthetic examples via ChatGPT: For the binary subtasks C2A and VIO (calls-to-action and violence glorification), we generated new positive examples to enrich the minority class. We prompted ChatGPT (GPT-4o) (OpenAI et al., 2024) in German to create 100 artificial training instances labeled TRUE for each task, instructing it to mimic the style and format of our data. The prompts used for data generation were:

- **VIO:** *„Könntest du bitte basierend auf den Trainingsbeispielen in der angehängten Datei 100 neue Trainingsbeispiele für die Klasse VIO = TRUE generieren? Die neuen Beispiele sollten denselben Aufbau wie die Originaldatei haben und im CSV-Format vorliegen.“*
- **C2A:** *„Könntest du bitte basierend auf den Trainingsbeispielen in der angehängten Datei 100 neue Trainingsbeispiele für die Klasse C2A = TRUE generieren? Die neuen Beispiele sollten denselben Aufbau wie die Originaldatei haben und im CSV-Format vorliegen.“*

The synthetic examples produced by ChatGPT were then reviewed for plausibility and correctness (to ensure they indeed reflected genuine calls-to-action or violence-glorifying content). Once vetted, we appended these 100 generated samples for VIO and 100 for C2A to the respective training sets. This approach of leveraging GPT-generated data is inspired by recent studies which showed that incorporating synthetic examples can boost performance on hate speech detection (Khullar et al., 2024; Schmidhuber and Kruschwitz, 2024; Girón et al., 2025). By expanding the positive class with diverse, machine-generated samples, we aimed to

expose the model to a wider variety of subtle harmful content phrases that were missing from the original data.

Manual addition of misclassified case: For the DBO task the model was sometimes too lenient. Some posts that were actually clear attacks on the democratic order (either agitational or subversive in nature) were wrongly classified by the model as mere "criticism." In other words, the model failed to recognize the extreme nature of these comments, treating them as if they were just normal policy critique. Through error analysis of the results on the test set, we identified 13 such misclassified instances.

These 13 posts contained strong anti-democratic sentiment (for example, calls to vote out all established parties or even references to a "revolution" with pitchforks), yet the model had placed them in the mild criticism category. To address this, we manually labeled each of these posts to its appropriate class (e.g. reassigning them from criticism to "agitation" or "subversive" as warranted) and added them to the DBO training set as additional examples. By including these cases, we reinforced the model's understanding of what truly constitutes an attack on the democratic order versus what is acceptable criticism, helping it draw a clearer line between normal dissent and dangerous subversive speech.

4.4 Training with Augmented Data

After augmenting the datasets as described, we retrained our models on the expanded training sets. All new examples (the 100 ChatGPT-generated positives for VIO, 100 for C2A, and 13 manually-added DBO criticism examples) were used alongside the original data in training. We then applied the same oversampling procedure on this combined set – oversampling the minority classes in the augmented training data. This ensured that the synthetic and newly added real examples were included in the oversampling, effectively amplifying their presence during training. The final models therefore learned from a balanced dataset that was both larger and more diverse than the original. We found this iterative enhancement of training data to be crucial: it allowed the models to correct some of their earlier mistakes by exposing them to informative new examples (both generated and hand-labeled) in the next training round.

4.5 Ensembling and Prediction

For each task, we fine-tuned multiple augmented models, as described above, and then used a soft-voting ensemble to make the final predictions. Each model in the ensemble outputs a probability distribution over classes, and we average these probabilities to determine the predicted label. This ensemble prediction was applied to the test set. The combination of data augmentation, oversampling, and ensembling yielded a noticeable improvement in capturing the minority classes without being overwhelmed by the majority class. In particular, augmenting with GPT-generated positives helped our classifiers detect subtle incitement and violence-glorifying cues they previously missed, and the additional criticism examples in DBO reduced false positives where the model had over-reacted to benign critiques.

Each team could submit a maximum of three runs. Our second run used the same ensemble configuration but without any data enrichment, relying solely on oversampling of the original training data. This variant achieved official F1-scores of 0.84 on C2A, 0.69 on DBO, and 0.79 on VIO. By comparison, our final ensemble with data enrichment obtained 0.83, 0.71, and 0.80 on the respective tasks. While the score for C2A showed a slight decrease of 0.01, detailed analysis of the predictions revealed the reason: between the two runs, 73 instances in the C2A test set changed predicted labels, with 34 switching from `FALSE` to `TRUE` and 39 from `TRUE` to `FALSE`. The enrichment thus helped to identify additional subtle calls to action but also led to some previously correct positive predictions being flipped to negative, which slightly reduced recall and thereby the overall F1-score. Table 2 shows representative examples of these shifts.

Overall, our methodology exemplifies an iterative refinement process: we start with a base trained model, analyze its weaknesses on the data, then address those weaknesses by expanding and rebalancing the training data before retraining. By doing so, we align our system with the nuanced nature of the GermEval 2025 tasks, ensuring that each classifier receives sufficient signal to recognize the rare but crucial harmful content, while not mislabeling innocuous content as dangerous.

5 Results and Discussion

The official leaderboard results for the GermEval 2025 shared task are presented in (Felser et al.,

Comment Text (translated excerpt)	Without Enrich.	With Enrich.
"... Dann wird gekämpft!" ("... then there will be fighting!")	FALSE	TRUE
"Schaut sie mal in Hamburg an." ("Just look at them in Hamburg.")	TRUE	FALSE
"Vielleicht die Leute von der Arbeitsagentur mal ARBEITEN lassen!" ("Maybe let the employment agency staff actually work!")	TRUE	FALSE

Table 2: Examples of C2A predictions that switched between the ensemble without data enrichment and the ensemble with data enrichment.

2025). Our system **nymera** delivered strong performance across all three subtasks: it ranked 3rd in the **C2A** task (Call-to-Action identification), 1st in **DBO** (detection of content attacking the free democratic basic order), and 4th in **VIO** (recognition of disturbingly positive statements towards violence). Notably, **nymera** achieved the *highest* score in the DBO subtask with an F1-score of 0.71, outperforming the second-place system by a margin of 0.02 . In the C2A task, **nymera** attained an F1-score of 0.83, narrowly behind the top two entries (0.87 and 0.84 by teams *SuperGLEBer* and *HSH;-)*, respectively). Similarly, for VIO our system reached 0.80 F1 , which, while competitive, trailed the best result (0.84) by 0.04. These outcomes confirm that **nymera** was among the leading systems in all categories, with an exceptional result in DBO and solid standings in C2A and VIO.

These results highlight the strengths of our approach. In particular, the first- place finish on DBO suggests that **nymera** excels at detecting content that targets the free democratic basic order. This strength may stem from the model's ability to capture domain-specific cues (e.g. terminology or hostile narratives) that characterize attacks on the free democratic basic order. Furthermore, **nymera**'s top-tier performance across all three tasks (ranking within the top four for each) demonstrates the system's robust generalization to different manifestations of harmful content. The architecture and training regimen we employed appear to transfer well between identifying overt incitement (C2A), anti-democratic rhetoric (DBO), and even subtle positive references to violence (VIO), underlining the versatility of the system.

Despite this strong overall performance, the areas where **nymera** did not secure the very top rank point to opportunities for improvement. In the C2A subtask, our system's third-place result indicates that at least two competing systems were better at recognizing calls to action. Similarly, the fourth-place outcome in VIO shows that a few systems surpassed **nymera** in detecting disturbingly positive or encouraging statements about violence. These relative gaps suggest that **nymera** may have missed certain linguistic cues or context nuances that are critical for these categories. For instance, calls to action can be implied indirectly or phrased in subtle ways (e.g., polite requests or coded language inciting action), and our model might not capture all such subtleties. In the case of VIO, distinguishing genuinely positive references to violence from sarcastic or context-dependent statements is challenging; the fact that other teams achieved higher scores implies that they may have incorporated more effective strategies for modeling such context or sarcasm. Thus, while **nymera** was competitive, missing the top-2 in C2A and VIO suggests room for refinement in handling the more nuanced or implicit instances of harmful content.

Several limitations of the current system could underlie these observations. One potential issue is overfitting. Our model was fine-tuned extensively on the provided training data for each subtask. This intensive fine-tuning, especially with limited training examples, raises the risk that **nymera** learned patterns too specific to the training set. If the model latched onto superficial cues that do not generalize (e.g., particular keywords or idiosyncratic patterns present in the training data), its performance on some unseen examples would be suboptimal. Although the system performed well on the official test sets, overfitting may still limit its robustness, particularly if deployed on content from different platforms or contexts not covered by the competition data.

Another limitation is the lack of deeper linguistic modeling in our approach. **nymera** relies on a transformer-based classifier without explicit incorporation of higher-level linguistic or world knowledge. It treats each input largely as a sequence of tokens, without modeling discourse-level information (such as conversation context or speaker intention) or leveraging linguistic features (such

as syntax or semantic roles). As a result, the system might not fully grasp cases where understanding pragmatics or context is necessary – for example, differentiating a genuine call to action from a figurative statement can require understanding the surrounding discourse or the speaker's intent. Similarly, recognizing a "positive" statement about violence may require understanding sarcasm or cultural references that go beyond surface text patterns. The absence of such enriched linguistic handling could have contributed to **nymera**'s slight underperformance in the more nuanced C2A and VIO tasks.

Additionally, our solution made use of synthetic data augmentation, which comes with its own trade-offs. We expanded the training set by generating or paraphrasing examples of harmful content to cover a wider range of expressions (especially for under-represented classes). While this strategy can improve recall and make the model more resilient to rare phrasings, it can also introduce distributional biases. If the synthetic examples are not perfectly reflective of real-world data, the model may learn artifacts specific to the artificially generated data. There is also a risk of the model over-relying on synthetic patterns at the expense of real data patterns. This reliance may have contributed to certain inconsistencies—for example, if some subtle forms of calls to action were underrepresented or inaccurately captured in the synthetic data, **nymera** might have struggled with those cases in the test set, while competing systems employing alternative strategies may have handled them more effectively.

6 Summary and Future Work

In summary, our unified system delivered strong, consistent performance across all three GermEval 2025 subtasks. It ranked 1st in the **DBO** task, 3rd in **C2A**, and 4th in **VIO**, confirming its effectiveness in detecting a diverse range of harmful content. These high rankings across the board underscore the versatility of our approach and its ability to generalize well to different manifestations of incitement, violent rhetoric, and anti-democratic speech.

This success can be attributed to several key strengths of our approach. First, the transformer-based ensemble (combining multiple fine-tuned models via soft-voting) enabled the system to capture complementary cues and improved its overall robustness. Second, extensive class balancing and data augmentation were critical: by oversampling minority classes and introducing synthetic examples, we mitigated severe class imbalance and exposed the models to a wider variety of rare but important patterns. Third, careful manual refinement—such as adding real misclassified examples for the DBO "criticism" class—helped correct systematic errors and sharpen the decision boundaries for under-represented categories. Together, these techniques enabled the system to reliably detect subtle forms of harmful content while minimizing false positives on benign inputs.

Looking ahead, we identify several avenues for further improvement. One direction is to explore a multi-task learning framework across the three subtasks, which could enable the model to leverage shared representations and cross-task signals (e.g., learning a general notion of "harmfulness" applicable to all categories). Another priority is to improve model calibration and decision threshold tuning to ensure optimal precision–recall trade-offs—especially important in high-stakes content moderation scenarios where false alarms and misses carry different costs. Finally, we aim to reduce reliance on synthetic data by incorporating additional real-world examples or external corpora into training. Expanding the training set with more authentic harmful content (drawn from varied sources) would help align the model with real-world distributions and mitigate any biases introduced by artificial examples. Pursuing these future directions should further enhance the robustness and generalizability of our system, paving the way for even more effective harmful content detection in German social media.

References

Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting calls to action in multimodal content: Analysis of the 2021 German federal election campaign on Instagram. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 1–13, Vienna, Austria. Association for Computational Linguistics.

Remi Calizzano, Malte Ostendorff, and Georg Rehm. 2021. DFKI SLT at GermEval 2021: Multilingual pre-training and data augmentation for the classification of toxicity in social media comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 25–31, Duesseldorf, Germany. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6788–6796. International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pages 4171–4186. Association for Computational Linguistics.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the GermEval 2025 Shared Task on Harmful Content Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. A transformer-based multitask framework for joint detection of aggression and hate on social media data. *Nat. Lang. Eng.*, 29(6):1495–1515.

Adrián Girón, Javier Huertas-Tato, and David Camacho. 2025. Llm synthetic generation to enhance online content moderation generalization in hate speech scenarios. *Computing*, 107:164.

Anna V. Glazkova, Michael Kadantsev, and Maksim Glazkov. 2021. Fine-tuning of pre-trained transformers for hate offensive and profane content detection in english and marathi. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 52–62. CEUR-WS.org.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Aman Khullar, Daniel Nkemelu, V. Cuong Nguyen, and Michael L. Best. 2024. Hate speech detection in limited data contexts using synthetic data generation. *ACM J. Comput. Sustain. Soc.*, 2(1).

Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.

Izzet Emre Kucukkaya and Cagri Toraman. 2025. Constructing ensembles for hate speech detection. *Natural Language Processing*, 31(3):745–770.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, volume 2826 of *CEUR Workshop Proceedings*, pages 87–111. CEUR-WS.org.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC subtrack at FIRE 2021: Hatespeech and offensive content identification in english and indo-aryan languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 1–19. CEUR-WS.org.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2024. Bert-based ensemble learning for multi-aspect hate speech detection. *Clust. Comput.*, 27(1):325–339.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 167–190. CEUR-WS.org.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, Shrey Satapara, Tithi Patel, and Hiren Madhu. 2022. Overview of the HASOC subtrack at FIRE 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language. In *Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022*, volume 3395 of *CEUR Workshop Proceedings*, pages 475–488. CEUR-WS.org.

Skye Morgan, Tharindu Ranasinghe, and Marcos Zampieri. 2021. WLV-RIT at GermEval 2021: Multitask learning with transformers to detect toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification*

of Toxic, Engaging, and Fact-Claiming Comments, pages 32–38, Duesseldorf, Germany. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahouti, Syed Ishtiaque Ahmed, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 72–84, Singapore. Association for Computational Linguistics.

Maximilian Schmidhuber and Udo Kruschwitz. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 - 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 352 – 363.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT & LT2 at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *CoRR*, abs/2004.11493.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. In Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand, editors, *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria - September 21, 2018*, pages 1 – 10.

Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Comput. Speech Lang.*, 89:101690.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

Yaqi Zhang, Viktor Hangya, and Alexander Fraser. 2024. A study of the class imbalance problem in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 38–51, Mexico City, Mexico. Association for Computational Linguistics.