abullardUR@GermEval Shared Task 2025: Fine-tuning ModernGBERT on Highly Imbalanced German Social Media for Harmful Content Detection

Samuel Bullard
Information Science Faculty
University of Regensburg
Regensburg, Germany
Samuel.Bullard@stud.uni-regensburg.de

Abstract

We present a system for the GermEval 2025 Shared Task on harmful content detection in German social media, addressing three tasks: Call to Action (C2A), Attacks on Democratic Basic Order (DBO), and Violence Detection (VIO). We fine-tune ModernGBERT on a dataset of German tweets from right-wing extremist networks (2014-2016) with severe class imbalance: 9.7% (C2A), 15.8% (DBO harmful), and 7.2% (VIO) positive ratios. Our experiments reveal that class-weighted cross-entropy loss outperforms the combined approach with Focal Loss, achieving macro-F1 scores of 0.82, 0.63, and 0.82 for C2A, DBO, and VIO respectively - improvements of +39%, +34%, and +19% over baselines. This suggests frequency imbalance, not hard example mining, is the primary challenge. While our models rank 4th/9 (C2A), 6th/9 (DBO), and 2nd/8 (VIO) on the leaderboard, nuanced categorization tasks requiring deep contextual understanding remain challenging, particularly for multi-class tasks.

1 Introduction

Harmful content detection is essential for online safety and democratic discourse in Germany. Three specific phenomena can pose societal risks: calls to action (C2A) can mobilize offline violence or criminal activity, attacks on democratic basic order (DBO) undermine constitutional principles and institutional legitimacy, while violence-related statements (VIO) normalize aggression and threaten individual safety. Each requires specialized detection approaches given their varied linguistic patterns and contextual markers.

This paper describes our system for the GermEval 2025 Shared Task (Communication Forensics Lab Mittweida, 2025; Felser et al., 2025), focusing on fine-tuning ModernGBERT with systematic loss function design for severe class imbalance. Our methodology centers on controlled comparison of class-weighted approaches and thorough dataset analysis to inform our modeling decisions.

Our contributions are threefold: (i) dataset characteristic analysis informing architectural choices, (ii) controlled comparison of class-weighted cross-entropy versus combined class-weights with focal loss, and (iii) competitive performance using a compact German encoder.

The papers' further structure is as follows: Section 2 surveys relevant work in the field of harmful content detection and class imbalance. Section 3 analyzes dataset properties that motivate our design choices. Section 4 details our system architecture and methodology. Section 5 presents results and discusses findings, limitations, and future directions. Section 6 concludes the paper.

2 Related Work

We summarize prior work on harmful content detection, German transformer models, and techniques for class imbalance that motivate our approach.

2.1 Early Approaches to Harmful Content Detection

Automated harmful-speech detection has progressed from rule-based filters to modern neural architectures. One of the earliest prototypes, *Smokey*, relied on manually crafted patterns to flag e-mail "flames" (Spertus, 1997). Traditional bag-of-words classifiers such as SVMs and logistic regression improved portability but still failed on creative orthography and implicit abuse (Davidson et al., 2017). Deep neural networks introduced convolutional and recurrent models that learn features directly from data; Collobert *et al.* demonstrated this paradigm on multiple NLP tasks and inspired early CNN-based hate-speech systems (Collobert et al., 2011). Nevertheless, limited context windows and vocabulary size constrained performance on nuanced, code-switched harmful speech (Kalchbrenner et al.,

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 320–326, 2025

2014; Magu and Luo, 2018).

2.2 Harmful Speech Detection in the German Social Media Context

In Germany the *Netzwerkdurchsetzungsgesetz* (NetzDG) mandates swift removal of illegal content, motivating research into reliable German-language detectors (Bundesministerium der Justiz und Bundesamt für Justiz, 2017). Initial resources were scarce: Ross et al. (2016) introduced one of the first publicly available German hate-speech corpora, a set of 541 refugee-related tweets. Community benchmarks filled this gap. The GermEval shared-task series supplies steadily larger corpora: the 2018 pilot on offensive tweets (Wiegand et al., 2019), its 2019 extension with implicit/explicit labels (Struß et al., 2019), and the 2021 task on toxic, engaging and fact-claiming comments (Risch et al., 2021) and finally the 2025 GermEval shared task on harmful-content detection (Communication Forensics Lab Mittweida, 2025). At the same time, specialised extremist resources emerged. Hartung et al. (2017) compiled a profile-level dataset of right-wing extremist Twitter users and identified symbolic codes such as "d18," a numeric reference to Adolf Hitler, among many others. The subsequent *Right-Wing Hate Speech Twitter corpus* (RWHN) enlarged coverage to 50k tweets and documents ciphers like "88" (standing for "Heil Hitler") alongside other evasive euphemisms (Jaki and Smedt, 2019). Together, research, legislation and shared tasks have contributed to driving a shift from lexicon-based heuristics to context-aware neural models in German NLP.

2.3 Transformer-Based Language Models for German

The highly parallelisable self-attention architecture of transformers has enabled BERT-style encoders to reach billions of parameters and to power moderation at an industrial scale. Flipkart's Unified BERT (Nayak and Garera, 2022) filters millions of user reviews for hate speech and policy violations in real time. Since self-attention allows every token to attend to every other token (Vaswani et al., 2017), BERT's bidirectional encoder (Devlin et al., 2019) captures the nuanced context required to detect implicit abuse.

Monolingual German transformers soon outperformed multilingual baselines: GBERT and its sibling GELECTRA (Chan et al., 2020) improved downstream accuracy on classification and sequence-tagging tasks. Yet the 512-token context limit inherited from BERT hinders the analysis of long snippets of text. Long-context adaptations addressed this gap. Longformer introduced sliding-window attention for inputs beyond 4000 tokens (Beltagy et al., 2020), while the ModernBERT architecture replaced absolute positions with rotary embeddings and paired them with sparse 128-token windows, scaling sequence length to 8192 tokens without quadratic memory growth (Warner et al., 2024). These advances are crucial for German content moderation, where a single post can now stretch to 25,000 characters on X (X Corp., 2025), exceeding the 512-token window and requiring long-range attention.

Building on this progress, *ModernGBERT* provides 134M to 1B-parameter German-only models trained on roughly 470 billion tokens and sets new state-of-the-art scores on German benchmarks (Ehrmanntraut et al., 2025). Large multilingual encoders such as EuroBERT narrow the monolingual advantage on general German natural language understanding (Boizard et al., 2025), yet on the public SUPERGLEBER Toxicity leaderboard (Pfister and Hotho, 2024) ModernGBERT-1B is among the strongest openly released encoder models for German hate-speech detection. Its compact 134M-parameter variant, which we fine-tune in our experiments, still achieves a competitive macro-F1 of 0.526 while requiring an order of magnitude fewer parameters.

2.4 Addressing Class Imbalance in Text Classification

Harmful-content corpora are oftentimes highly skewed. Offensive or toxic instances constitute roughly one third of the tweets in GERMEVAL 2018 (34%) and 2019 (32%), and 35% of the Facebook comments in the GERMEVAL2021 toxic-comment task (Wiegand et al., 2019; Struß et al., 2019; Risch et al., 2021). For the 2025 GERMEVAL shared task we measured 9.7% Call-to-Action, 15.8% Democratic Basic Order and 7.2% Violence in the released training data.

Such imbalance biases models toward the majority class, so recent shared tasks rank systems by macro-F1 instead of accuracy (Sokolova and Lapalme, 2009; Risch et al., 2021).

Mitigation methods fall into three families:

1. **Data-centric re-balancing.** Random over- or undersampling, SMOTE (Chawla et al.,

2002) and its variants, back-translation or lexicon-guided augmentation (for example *AbusiveLexiconAug*) alter the training distribution (Krawczyk, 2016; Fernández et al., 2018; Zhang et al., 2024).

2. **Loss-centric re-weighting and margin tuning.** Inverse-frequency class weights are a common baseline; the winning *hpi-DEDIS* run at GermEval 2019 used weighted cross-entropy with weights {6.57, 1.96, 1.56, 0.37} for PROFANITY, ABUSE, INSULT and OTHER, respectively (Risch et al., 2019). Focal Loss (Lin et al., 2017), Label-Distribution-Aware Margin (LDAM) loss (Cao et al., 2019), Class-Balanced Loss (Cui et al., 2019) and Logit-Adjusted Loss (Menon et al., 2021) further reduce majority bias.

3. **Prediction-centric calibration and ensembling.** Threshold moving, cost-sensitive bagging or boosting, and majority-vote ensembles (used by several GermEval 2021 teams (Bornheim et al., 2021; Risch et al., 2021)) correct residual bias during or immediately after training.

A cross-corpus study covering eight abusive-language datasets finds weighted cross-entropy and Focal Loss to be the most robust single-model choices (Zhang et al., 2024). We therefore benchmark both losses within the ModernGBERT framework and the confines of the GermEval 2025 shared task.

3 Dataset Analysis

This section summarizes dataset properties that inform our modeling choices. The class distributions for the three subtasks are shown in Table 1.

3.1 Basic Dataset Statistics

3.2 Text Length Characteristics

Tweets are short on average but with long outliers. Approximately 75% of posts are under 180 characters, yet outliers up to 10k characters motivate a long-context encoder to avoid truncation.

3.3 Cross-Dataset Overlap

Substantial content overlap exists across subtasks (e.g., >60% of unique descriptions appearing in at least two subtasks), reinforcing the potential of shared representations and the need for careful split design to avoid leakage when experimenting with joint models and splitting datasets into train/validation/(test) splits.

Task / Class	Count	Share
C2A		
False	6,177	90.3%
True	663	9.7%
DBO		
Nothing	6,277	84.2%
Criticism	804	10.8%
Agitation	313	4.2%
Subversive	60	0.8%
VIO		
False	7,219	92.8%
True	564	7.2%

Table 1: Class distributions in the training data.

3.4 Linguistic Characteristics

German compounds, social-media artifacts (hashtags, mentions, URLs), and coded language (e.g., numeric ciphers) occur frequently and influence our model choice and preprocessing rationale.

4 System Architecture and Methodology

4.1 Model Architecture

We fine-tune ModernGBERT-134M (Ehrmanntraut et al., 2025; Warner et al., 2024) for each subtask with a mean-pooled classification head. ModernGBERT's 8192 token context length accommodates the vast majority of lengths in our dataset. We train separate models for each subtask end-to-end.

4.2 Loss Design for Class Imbalance

The dataset is severely imbalanced across all subtasks (see Section 3). We implement two strategies:

4.2.1 Class-Weighted Cross-Entropy Loss

We apply inverse frequency weighting to the cross-entropy loss function (King and Zeng, 2001):

$$w_i = \frac{n}{n_i \cdot k} \tag{1}$$

where w_i is the weight for class i, n is the total number of samples, n_i is the number of samples in class i, and k is the number of classes. The weighted loss becomes:

$$\mathcal{L}_{WCE} = -w_y \log(p_y) \tag{2}$$

This weighting ensures that mistakes on minority classes incur proportionally higher penalties during training.

4.2.2 Combined Approach: Class Weights with Focal Loss

In our second approach, we combine class weighting with Focal Loss (Lin et al., 2017) to address both frequency imbalance and hard example mining:

$$\mathcal{L}_{FL} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (3)$$

where p_t is the predicted probability for the true class, α_t is the class-specific weight (equivalent to w_y above), and γ is the focusing parameter. The term $(1-p_t)^{\gamma}$ acts as a difficulty multiplier: when the model is confident and correct ($p_t \approx 1$), this term approaches zero, effectively down-weighting easy examples. Conversely, for hard examples where the model assigns low probability to the correct class, this term remains significant.

We optimize γ in the range $[0.5, 3.0]$ using Bayesian hyperparameter optimization (Snoek et al., 2012) via Weights & Biases (Biewald, 2020) sweeps.

4.3 Preprocessing and Tokenization

The dataset undergoes minimal preprocessing to preserve the authentic characteristics of social media text and leverage the contextual understanding capabilities of transformer models:

- **Existing anonymization tokens** from the dataset are retained: `[@POL]` (police/police authorities), `[@GRP]` (groups/organisations/associations), `[@IND]` (individuals), and `[@PRE]` (press/press offices/news portals)
- **Additional anonymization** is performed: URLs are replaced with `[@URL]` and email addresses with `[@EMAIL]` using the `cleantext` library
- **Data leakage prevention:** We remove validation samples with exact text duplicates from the training set to prevent inflated validation metrics, thereby misguiding the training process. This removed 2.7% (C2A), 6.1% (DBO), and 4.6% (VIO) of validation samples.
- Text is tokenized using ModernGBERT's tokenizer.
- No additional normalization, lowercasing, or cleaning is applied to maintain the original linguistic characteristics

We preserve anonymization tokens as they carry contextual information - e.g., `[@POL]` references may be more relevant for detecting attacks on democratic order than `[@IND]` references.

4.4 Training Configuration

We optimize hyperparameters using Bayesian search over learning rate $\{1\times10^{-5}, 2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}\}$ (discrete), weight decay $[0.01, 0.1]$ (continuous), batch size $\{8, 16\}$ (discrete), warmup steps $\{100, 300, 500\}$ (discrete), and focal loss gamma $[0.5, 3.0]$ (continuous). Models use AdamW with linear scheduling, early stopping (patience=5), and are selected based on validation macro-F1.

4.5 Evaluation Metrics

We use macro-F1 as the primary evaluation metric (as used by the shared task leaderboard on the hidden test set). On the validation split we report macro-F1, accuracy, macro-precision, macro-recall, and Matthews correlation (MCC). For binary tasks we additionally report ROC-AUC and Average Precision, and we include per-class precision/recall/F1 and support. Please note, that these validation-based metrics are optimistically biased since they steered the training process. For C2A and VIO, we will also report zero-shot results on the external GERMEVAL-2018 (coarse) test set. These per-model metrics will be provided in the GitHub repository (see Section 6). Metrics are reported on the validation split, as the hidden test set for the GermEval2025 shared task was not publicly released by the organizers, post-competition.

5 Results and Discussion

This section first presents the overall results and loss-design comparison, then discusses key findings and limitations.

5.1 Results

Main Results. Our main results on the competition organizers' hidden test set are presented in Table 2.

Leaderboard. We achieved 4/9 (C2A), 6/9 (DBO), and 2/8 (VIO) by macro-F1.

Hyperparameter Search and Loss Comparison. We conducted limited Bayesian hyperparameter sweeps (10 iterations per subtask) using Weights & Biases (Biewald, 2020). Full sweep configurations

Method	C2A	DBO	VIO
Official Baselines			
Gradient Boosting + SBERT	0.59	–	–
Cost-Sensitive Linear SVM	–	0.47	–
Qwen2.5 LLM Few-Shot	–	–	0.69
Our Approaches			
ModernGBERT + Class Weights	**0.82**	**0.63**	**0.82**
ModernGBERT + Weights + Focal	0.82	0.56	0.81

Table 2: macro-F1 on hidden test set data. Official baseline per subtask from the shared task overview (Felser et al., 2025).

and run artifacts will be provided alongside the code and models. Class-weighted cross-entropy consistently matched or outperformed the combined focal-loss variant across subtasks.

5.2 Discussion

Key Findings. Class-weighted cross-entropy is sufficient for this dataset and choice of model. Focal loss did not improve performance and seems to down-weight scarce correctly classified minority examples.

Limitations and Future Directions. Results derive from a single dataset and time period, which makes generalization to other domains less effective. Future work should include methods of data augmentation for extreme minorities and exploring multi-task learning, given the substantial overlap across tweet descriptions in all three subtasks.

6 Conclusion

We presented a system for the GermEval 2025 shared task, focusing on handling the severe class imbalance present in the data. Our primary finding is that class-weighted cross-entropy is more effective than a combined approach with Focal Loss. Our results suggest that more complex, combined techniques are not always superior for real-world NLP tasks. The success of class weighting over Focal Loss demonstrates the importance of understanding problem-specific characteristics, dataset characteristics and the likely impact a specific technique has on the performance of a model rather than applying state-of-the-art methods uncritically.

For practitioners deploying harmful content detection systems, our findings indicate that ModernGBERT with appropriate class weighting can achieve competitive performance on German social media text. However, the 63% macro-F1 on the nuanced DBO task underscores that automated systems should augment rather than replace human moderation, particularly for legally sensitive categorizations requiring contextual understanding.

Acknowledgments

I thank Prof. Dr. Udo Kruschwitz for supervision and guidance throughout this research. I also acknowledge the GermEval 2025 organizers from the Communication Forensics Lab at Mittweida University of Applied Sciences for providing the dataset and organizing the shared task.

Code, Metrics and Model Availability

The code, trained models, evaluation metrics, sweep artifacts, and detailed analyses will be made available at: `https://github.com/abullard1/abullardUR-GermEval-Shared-Task-2025`

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 152–168, Online. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, and 1 others. 2025. Eurobert: Scaling multilingual encoders for european languages. *Preprint*, arXiv:2503.05500.

Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2021. FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 105–111, Duesseldorf, Germany. Association for Computational Linguistics.

Bundesministerium der Justiz und Bundesamt für Justiz. 2017. Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz – netzdg). Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 61, S. 3352. Ausfertigungsdatum 1. September 2017, verkündet am 7. September 2017.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Preprint*, arXiv:1906.07413.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational*

Linguistics, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Preprint*, arXiv:1103.0398.

Communication Forensics Lab Mittweida. 2025. Germeval 2025 shared task: Harmful content detection. Organised by the Communication Forensics Lab, Mittweida.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. *Preprint*, arXiv:1901.05555.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. Moderngbert: German-only 1b encoder model trained from scratch. *Preprint*, arXiv:2505.13136.

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.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. Learning from imbalanced data sets. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):657–675.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–33, Copenhagen, Denmark. Association for Computational Linguistics.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. *CoRR*, abs/1910.07518.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *Preprint*, arXiv:1404.2188.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9:137–163.

Bartosz Krawczyk. 2016. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988. IEEE.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. *Preprint*, arXiv:2007.07314.

Ravindra Nayak and Nikesh Garera. 2022. Deploying unified BERT moderation model for E-commerce reviews. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 540–547, Abu Dhabi, UAE. Association for Computational Linguistics.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics.

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, Duesseldorf, Germany. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *Conference on Natural Language Processing*, pages 251–257.

Bjoern Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. Preprint. University of Duisburg-Essen repository.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Preprint*, arXiv:1206.2944.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.

Ellen Spertus. 1997. Smokey: automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, page 1058–1065. AAAI Press.

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 German Society for Computational Linguistics, editor, *Proceedings of the 15th Conference on Natural Language Processing (KONVENS) 2019*, pages 354–365. s.a., Nürnberg/Erlangen.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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. Austrian Academy of Sciences Press.

X Corp. 2025. About different types of posts. X Help Center. Accessed 20 Jul 2025.

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.