Hate speech detection in German twitter comments

Silvio Klenk

11th March 2025

Abstract

This paper presents a comprehensive analysis of German hate speech detection using the *DeTox* dataset. Four methods are compared: two transformer-based models (one domain-specific, one general-purpose), a large language model (ChatGPT), and a classical TF-IDF + XGBoost baseline. The domain-specific BERT model attains the highest accuracy, while ChatGPT exhibits strong recall but generates more false positives. The general-purpose `german-sentiment-bert` model offers a balanced performance, though it trails the general-purpose model in raw metrics. Subtle or implicit hate speech remains a persistent challenge, especially for group-targeted remarks and coded references. These findings underscore the complexity of hateful language in German and the potential benefits of domain-adaptive fine-tuning, interpretability, and context-aware modeling for more effective content moderation. [1]

[1] I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and

Contents

List of Figures

List of Tables

acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are serious and grave offenses in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion, or copying. This assignment, or any part of it, has not been submitted previously by me or any other person for assessment on this or any other course of study.

1 Introduction

Hate speech detection is an essential task in the moderating of online platforms, as harmful discourse can lead to social polarization and real-world violence. Although English-language resources are plentiful, German-language research has been more limited. The *DeTox* dataset [4] addresses this gap by offering a substantial collection of German comments with fine-grained annotations, including explicit and implicit hate speech indicators, target attributes, and multiple discrimination types.

Advances in transformer-based language models [2] have led to impressive performance gains on various natural language processing tasks, including hate speech detection. Large language models (LLMs), such as GPT-3 [1], further extend these capabilities through few-shot and zero-shot learning. However, empirical evaluations in German often require domain-specific adaptations. For instance, `german-sentiment-bert` [3] has been trained on extensive German corpora and may capture linguistic nuances better than general-purpose models.

This paper examines multiple approaches for German hate speech detection on the DeTox dataset: two transformer models (one general purpose, one domain specific), a large language model (ChatGPT), and a classical TF-IDF + XGBoost baseline. A uniform experimental setup is applied to compare their abilities to handle subtle or implicit hateful content, diverse target categories, and various discrimination dimensions. The following sections detail the dataset, methodology, experimental results, and key insights into model performance and limitations.

2 Research Question and Methodology

This section defines the task of detecting hate speech in German social networks and outlines the approach adopted to train and evaluate multiple models in the DeTox dataset [4].

2.1 Problem Definition

Hate speech detection is formalized here as a binary classification task, where each comment is labeled as hateful or non-hateful. In line with prior work [4], a threshold of 0.33 is applied to the continuous *hate_speech* score in DeTox to obtain a binary label. The objective is to assess whether a model can correctly identify hateful content, including subtle or implicit remarks that may not be overtly abusive.

The DeTox dataset encompasses a wide range of annotations, such as:

- **Explicit vs. Implicit Expression:** Indicators of whether a comment is openly hateful or context-dependent.
- **Target Categories:** Person, group, or public entities.
- **Discrimination Types:** Attributes such as `discrim_nation` or `discrim_religion`, `discrim_sexIdent`, `discrim_socialStatus`, `discrim_Ethnicity`, denoting prejudicial remarks.

These dimensions highlight the complexity of the problem, as language can be ambiguous or coded in ways that elude simple keyword-based approaches. The subliminal and widespread behavior of hate speech should also be understood as problematic in terms of recognition. Some characteristics can be attacked without the direct use of hate speech, but simply by dragging the characteristics to the ground.

2.2 Proposed Approach

A unified experimental framework is established to compare four distinct methods:

1. **Transformer Model 1:** A domain-specific German BERT variant (`german-sentiment-bert` [3]) fine-tuned for binary classification.
2. **Transformer Model 2:** A general-purpose German BERT model [2], also adapted to detect hate speech through a new linear classification head.
3. **ChatGPT:** A large language model [1] accessed via prompt engineering in German, returning `1` for hateful content and `0` otherwise.
4. **TF-IDF + XGBoost:** A baseline where the cleaned text is vectorized using TF-IDF, and an XGBoost classifier is trained to discriminate hateful vs. nonhateful remarks.

Each model is trained or prompted under consistent conditions. The data set is divided into a 80% training subset and a 20% validation subset, stratified by the binary label. This ensures a fair comparison of performance metrics across models. During preprocessing, comments are lowercased, punctuation is removed, and lemmatization is optionally applied. For implicit remarks or subtle discrimination cues, the annotated attributes of the data set are used in subsequent error analyzes rather than as direct input features.

2.3 Evaluation Protocol

A stratified 20% validation set is used for final metrics:

- **Accuracy, Precision, Recall, F1:** Standard metrics are reported to capture both the overall correctness and the balance between false positives and false negatives.
- **Error Analysis:** Misclassifications are examined in explicit / implicit columns, target categories, and discrimination attributes. This reveals whether certain forms of hateful content systematically elude detection.
- **Aggregated Predictions:** All model predictions are merged on a per-comment basis (via *c_text*) to facilitate cross-model comparisons and identify overlaps in failure cases.

The next section presents empirical findings, highlighting each model's performance and exploring where each approach succeeds or fails in capturing hate speech under varying linguistic and contextual conditions.

3 Experimental Results

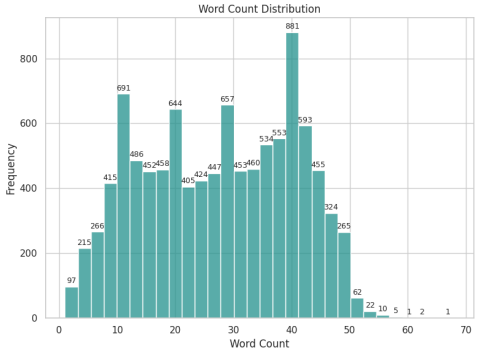
This section presents empirical findings from four perspectives: (1) exploratory data analysis (EDA), (2) final performance metrics, (3) model training and prediction curves, and (4) a detailed error analysis. All results reflect the 20% validation subset described in Section 2.

3.0.1 EDA Insights

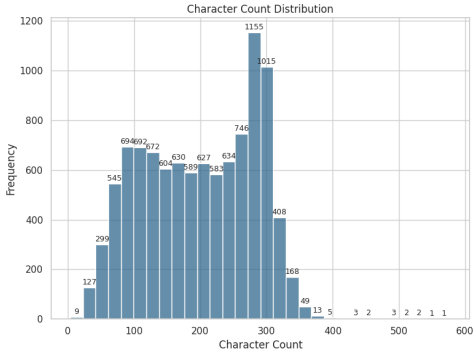
A preliminary examination focused on text length, term frequencies, and token embeddings. Figure 1 combines word and character count distributions, suggesting moderate-length comments but also some outliers. Figure 2 displays the top 20 words, bigrams, and a word cloud, while Figure 3 uses t-SNE to cluster the top 100 terms, indicating thematic groupings (e.g., public figures, hateful slurs).

Most entries fall in a moderate range, though some are very short or quite long, as shown in 1.

2 contains a visible representation of commonly used words in all documents. Personal names and certain hateful triggers appear often, reflecting the emphasis on social media commentary. Visualization of embeddings for the top 100 tokens, suggesting clusters

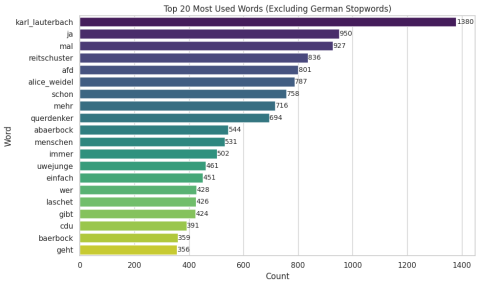


(a) Word count distribution

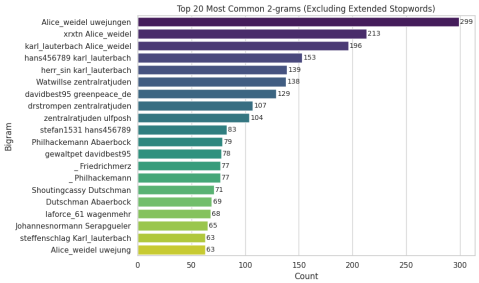


(b) Character count distribution

Figure 1: Comment length analysis



(a) Top 20 words



(b) Top 20 bigrams

(c) Word cloud

Figure 2: Frequent terms in the dataset

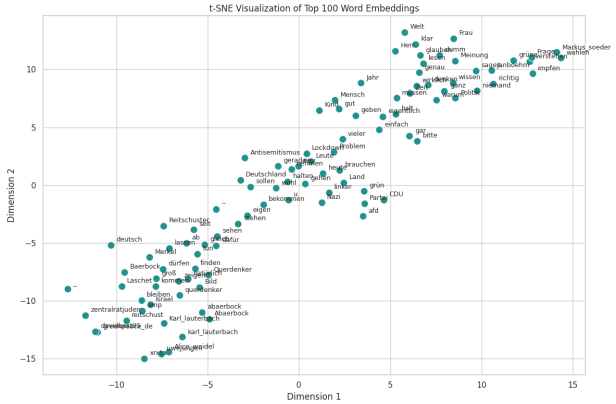


Figure 3: t-SNE visualization

(e.g., personal references vs. general hate terms) 3. Observations from these EDA plots confirm that the dataset contains references to public figures and hateful language patterns. A considerable portion of terms relates to socio-political discourse, aligning with the need for robust detection methods.

3.0.2 Final Performance Metrics

Table 1 shows the accuracy, precision, recall, and F1 scores for the four approaches: TF-IDF + XGBoost, ChatGPT, Transformer Model 1 (`german-sentiment-bert`), and Transformer Model 2 (general-purpose German BERT). The results reflect the final aggregator approach discussed earlier.

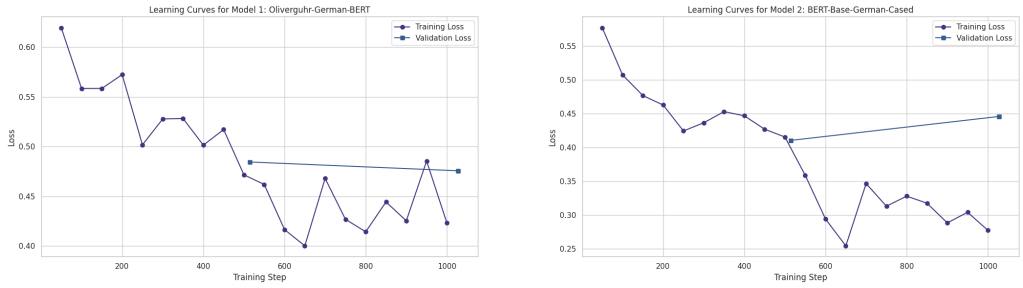
Table 1: Final metrics on the DeTox validation set.

Model	Accuracy	Precision	Recall	F1
TF-IDF + XGBoost	0.76	0.63	0.30	0.41
ChatGPT	0.66	0.47	0.86	0.61
Transformer Model 1	0.82	0.70	0.58	0.63
Transformer Model 2	0.79	0.65	0.50	0.56

The Domain-specific Transformer Model 1 achieves the highest accuracy, whereas ChatGPT demonstrates superior recall but lower precision, often labeling borderline content as hateful. TF-IDF + XGBoost performs respectably in overall accuracy but fails to capture many implicit or subtle cues, reflected in the low recall. general-purpose Model 2 (`german-sentiment-bert`) maintains a balanced performance but slightly trails the Domain-specific model in raw metrics.

3.0.3 Model Training and Prediction Curves

Figures 4 and 5–6 illustrate training convergence, ROC curves, and precision-recall curves for both transformer models.



(a) Learning curve: Transformer Model 1

(b) Learning curve: Transformer Model 2

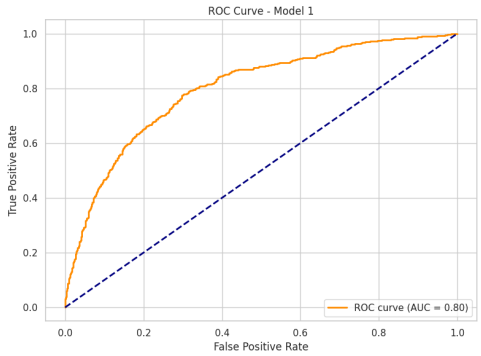
Figure 4: Training vs. validation loss over epochs

Figure 5 shows the ROC and precision-recall curves for Model 1, while Figure 6 does the same for Model 2. Both curves confirm moderate to strong separability, with Model 2 showing a slightly higher area under the ROC curve (AUC). In terms of precision-recall, Model 2 also exhibits a better balance, aligning with the final metrics in Table 1.

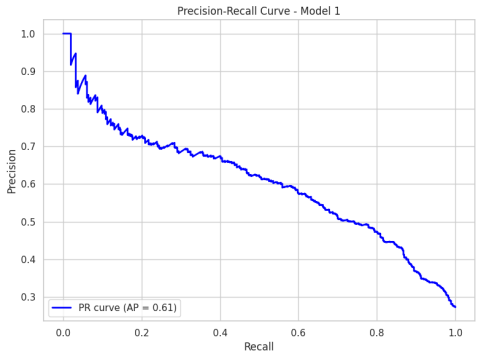
3.0.4 Error Analysis

A merged DataFrame of all predictions facilitates cross-model comparisons. Figure 7 plots how many comments each model misclassifies (0 to 4). A notable cluster of comments is misclassified by multiple models, often those with implicit or group-targeted content.

Further breakdowns highlight specific difficulties:

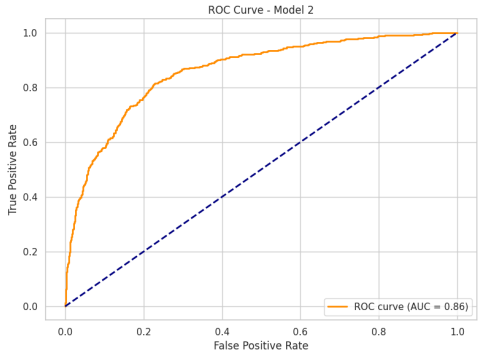


(a) ROC curve (Model 1)

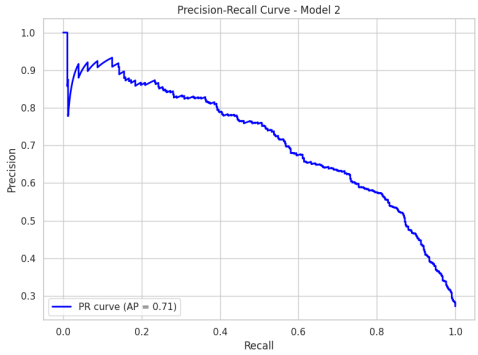


(b) Precision-Recall curve (Model 1)

Figure 5: Model 1 (`german-sentiment-bert`) curves



(a) ROC curve (Model 2)



(b) Precision-Recall curve (Model 2)

Figure 6: Model 2 (general-purpose German BERT) curves

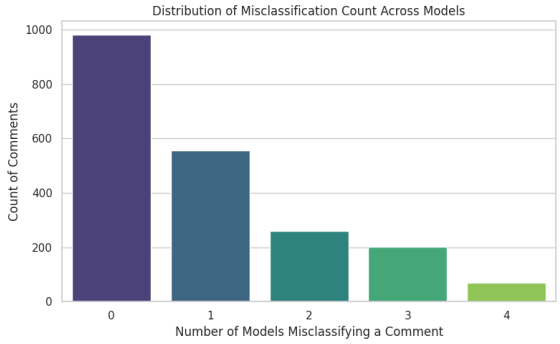
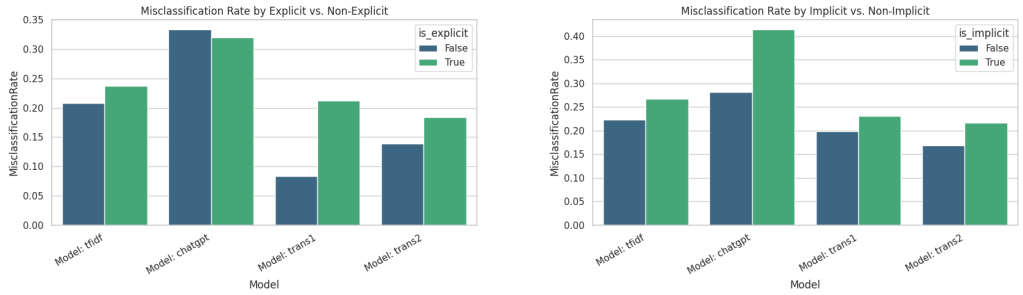


Figure 7: Distribution of misclassification counts across four models

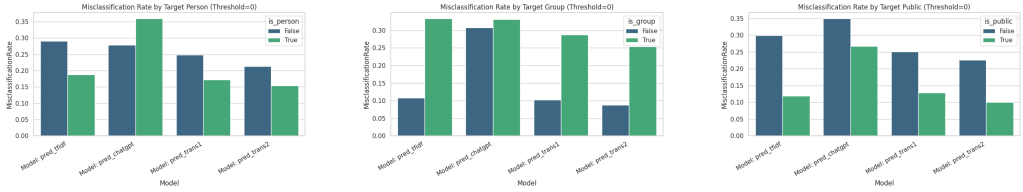
- **Explicit vs. Non-Explicit:** Figure 8(a) indicates that explicit hate speech is somewhat easier for TF-IDF and transformers to catch, while ChatGPT occasionally over-predicts hate speech, leading to moderate improvements on explicit content.
- **Implicit vs. Non-Implicit:** Figure 8(b) shows higher misclassification rates for implicit remarks across all models, consistent with earlier observations.
- **Target Types (Person, Group, Public):** Figures 9 (a)–(c) confirm that group-targeted remarks pose a greater challenge, aligning with the increased complexity in identifying group-based hateful language.

- **Discrimination Types:** Figure 10 compares error rates when `discrim_nation` or `discrim_religion` are flagged, revealing large positive spikes for TF-IDF and moderate increases for transformers, whereas ChatGPT sometimes shows fewer errors but higher false positives overall.



(a) Explicit vs. Non-Explicit (b) Implicit vs. Non-Implicit

Figure 8: Comparison of average misclassification rates for explicit/implicit content



(a) Person (b) Group (c) Public

Figure 9: Misclassification rates based on target category

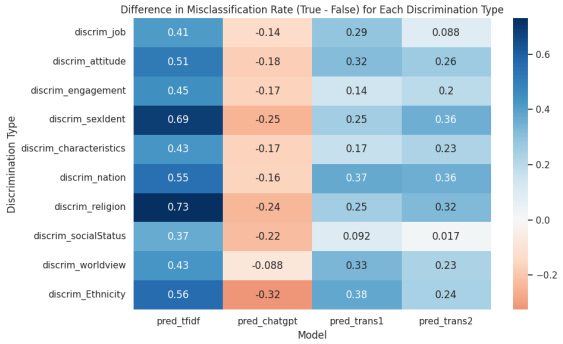


Figure 10: Differences in misclassification rates when certain discrimination types are present

These observations emphasize the challenges of subtle language cues and group-based hostility, as well as the tendency for ChatGPT to produce higher recall with increased false positives. The domain-specific BERT model demonstrates moderate success in capturing certain coded language, but general-purpose BERT still outperforms it in raw metrics, underscoring the complexity of German hate speech detection. The final section offers concluding remarks and future directions.

4 Concluding Remarks

This paper investigated German hate speech detection on the *DeTox* dataset [4], comparing two transformer-based models, ChatGPT, and a TF-IDF + XGBoost baseline.

Experimental results demonstrated that a general-purpose German BERT model consistently achieved the highest accuracy, while ChatGPT showed superior recall but at the cost of more false positives. The domain-specific `german-sentiment-bert` model offered a balanced approach, though it slightly trailed the general-purpose variant in raw metrics. TF-IDF + XGBoost performed competitively on overt content but struggled to capture implicit cues, resulting in lower recall.

Error analyses highlighted the persistent challenges posed by implicit language, group-based hostility, and coded references, all of which can evade simple keyword matching. Group-targeted hate speech and discrimination types such as `discrim_nation` or `discrim_religion` frequently correlated with higher misclassification rates, underscoring the complexity of subtle or context-dependent expressions of hostility. Although ChatGPT often detected coded content more aggressively, it occasionally produced spurious hate predictions, reducing precision.

Looking forward, several directions may enhance performance and robustness:

- **Contextual Modeling:** Incorporating conversation history or thread-level metadata could clarify ambiguous remarks and improve detection of subtle hate speech.
- **Domain Adaptation:** Further fine-tuning on specialized subsets of German social media data, especially for underrepresented discrimination categories, might increase recall without excessive false positives.
- **Interpretability and Feedback Loops:** Techniques like SHAP or LIME can elucidate model decisions, while user or moderator feedback can refine boundary cases in real-time moderation settings.

Overall, the findings confirm that hate speech detection in German demands both robust linguistic coverage and sensitivity to nuanced or implicit cues. Continuous expansion of domain-specific resources, along with context-aware modeling and interpretability, stands to reduce error rates and improve the reliability of automated content moderation systems.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[3] Oliver Guhr. Training a broad-coverage german sentiment classification model using fine-tuned BERT. arXiv preprint arXiv:2005.09174, 2020. `https://arxiv.org/abs/2005.09174`.

[4] Johannes Wich, Christian Schwiegelshohn, and Ulf Brefeld. Do we need fully-labeled data for hate speech detection? the case of the german DeTox dataset. In Christian Hammer, Jörg Hoffmann, Stefan Hartmann, and Heike Wehrheim, editors, *KI 2020: Advances in Artificial Intelligence*, volume 12325 of *Lecture Notes in Computer Science*, pages 28–40. Springer, 2020.