A Novel Dataset for Classifying German Hate Speech Comments with Criminal Relevance

Vincent Kums[1], Florian Meyer[2], Luisa Emily Pivit[3], Uliana Vedenina[3], Jonas Wortmann[3], Melanie Siegel[3], Dirk Labudde [1,2]

[1]Hochschule Fresenius
[2]Hochschule Mittweida - University of Applied Sciences
[3]Hochschule Darmstadt - University of Applied Sciences
Correspondence: florian.meyer@hs-mittweida.de

Abstract

The consistently high prevalence of hate speech on the Internet continues to pose significant social and individual challenges. Given the centrality of social networks in public discourse, automating the identification of criminally relevant content is a pressing challenge. This study addresses the challenge of developing an automated system that is capable of classifying online comments in a criminal justice context and categorising them into relevant sections of the criminal code. Not only technical, but also ethical and legal requirements must be considered. To this end, 351 comments were annotated by public prosecutors from the Central Office for Combating Internet and Computer Crime (ZIT) according to previously formed paragraph classes. These groupings consist of several German criminal law statutes that most hate comments violate. In the subsequent phase of the research, a further 839 records were assigned to the classes by student annotators who had been trained previously.

1 Introduction

The number of hate comments reported on social media continues to increase, as confirmed by the latest statistics from the German Federal Criminal Police Office (BKA) (Bundeskriminalamt, 2024). The European Union's Digital Services Act aims to protect users from insults, threats, and harassment by requiring platform providers to review and, if necessary, delete reported content within a specified timeframe. In addition, platforms must report offensive content to authorities, who evaluate its legal relevance for possible prosecution (Digital Service Act, 2022). However, due to the high volume of cases, courts and prosecutors are often overwhelmed (HessenGegenHetze, 2025). Developing reliable automated methods for legally classifying digital thus remains a central challenge for authorities. These methods must strike a balance between technical capabilities and the legal requirement for a robust and reliable dataset, which this paper provides. Despite efforts to define clear boundaries between legal and illegal content, practical application often reveals ambiguity, with seemingly similar statements judged differently depending on context and interpretation. The dataset presented here serves as a foundation for further research on automated classification of potentially criminal content. While these tools can aid in identifying such content, the final judgment ultimately rests with the courts.

The paper contains examples of hate speech to illustrate the issues. The authors explicitly disagree with these examples and use them solely for analytical purposes.

2 Related Work

2.1 Motivation

Many hate speech detection datasets focus primarily on broad classifications such as sentiment, toxicity or discrimination (Bertram et al., 2023; Jahan and Oussalah, 2023). While these are valuable for content moderation, they lack the granularity required for precise legal assessment. A dataset explicitly aligned with legal definitions would enable the development of more accurate classification models, distinguishing lawful from unlawful speech using established legal standards rather than subjective or community guidelines. This enables legal experts and law enforcement to evaluate online discourse within a well-defined legal framework.

Hate speech often does not fit into a single legal category, but can violate multiple legal provisions simultaneously. Traditional single-label or binary classification approaches fail to capture this complexity. As legal decisions made by the relevant authorities depend on the interpretation of specific legal statutes, each class in our dataset corresponds to the relevant sections of the Strafgesetzbuch (StGB),

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 41–52
August 1, 2025 ©2025 Association for Computational Linguistics

ensuring alignment with legal assessments.

We present a new multi-label dataset of illegal and legal hate comments compiled from various sources. Each comment is assigned to one of several categories, indicating whether:

- it is presumably legal
- it presumably constitutes **one** offense under the German Criminal Code (referred to as StGB), e.g., only disseminates unconstitutional material,
- it presumably constitutes **multiple** offenses under the StGB, e.g., disseminates unconstitutional material and calls for public violence.

2.2 Related Datasets

Several annotated hate speech datasets have been introduced to advance research in this area. Although most of these datasets are in English, such as HateXplain (Mathew et al., 2020) and AbuseEval v1.0 (Caselli et al., 2020), the availability of German hate speech datasets has only recently started to increase. The first German dataset, published by (Ross et al., 2016), contained around 500 tweets with binary annotations. Shared tasks such as GermEval and HASOC have contributed significantly to this field by providing multiple datasets focused on German hate speech (Wiegand and Siegel, 2018; Struß et al., 2019; Mandl et al., 2019). The DeTox dataset (Demus et al., 2022) is notable for including conversation threads, placing individual comments within their broader context. Most of these datasets, including those mentioned above, draw data primarily from social media platforms such as Twitter (now X) and Facebook. The most recent addition is the GAHD (German Adversarial Hate Speech Dataset) (Goldzycher et al., 2024), which includes 11,000 adversarial hate speech in German language. Although some datasets focus on specific targets of hate speech, such as offensive comments against foreigners (Bretschneider and Peters, 2017) and refugees (Ross et al., 2016), research on the legal aspects of hate speech in the German context remains limited. One of the few works in this area is by (Zufall et al., 2019), which examines the automated classification of political Twitter posts under three sections of German law. Furthermore, (Schäfer, 2023) introduced a data set designed to detect potentially illegal hate speech, explicitly covering five sections of the German Criminal Code (StGB). The DeTox dataset under discussion also included annotations regarding the criminal relevance of the posts under German criminal law. However, the authors noted that these annotations were made without any legal background. While these studies provide valuable insights into the detection of legally relevant hate speech, they are limited in terms of the range of crimes covered and the diversity of annotation sources. Building on this foundation, our work presents a new dataset that extends previous efforts by covering nine German criminal laws related to hate speech, providing a more comprehensive classification framework. A key distinction of our dataset lies in its dual annotation process: public prosecutors ensure high legal accuracy, while additional comments are annotated by trained student annotators and a professor, divided into two groups. This combined approach allows for a more nuanced assessment of criminal hate speech, addressing both the need for expert legal perspectives and the challenges of scalability in machine learning.

2.3 Definition of Hate Speech

Although the concept of hate speech is widely discussed in academic literature and political frameworks, there is no universally accepted definition. Institutions such as the European Commission against Racism and Intolerance (ECRI), the United Nations (UN), and major social media companies such as Meta offer slightly different definitions, typically emphasizing discriminatory or inflammatory speech based on protected characteristics (European Commission against Racism and Intolerance (ECRI), 2024; United Nations (UN), 2024; Meta, 2024). However, these definitions serve primarily as ethical or community guidelines, rather than legally binding standards. Unlike several countries with explicit laws against online hate speech (Strafgesetzbuch, 2024), German criminal law does not have a separate provision specifically criminalizing online hate speech. Instead, relevant cases are prosecuted under existing sections of the German Criminal Code (StGB), such as Section 185 (Insult) to protect personal rights, Section 111 (Public Incitation to Crime) to prevent incitement to violence, and Section 130 (Incitement of the People) for protecting against violence or hate based on nationality, ethnicity, religion, or other identity factors. This highlights the difficulty in systematically classifying hate speech with criminal relevance, as the legal assessment depends not only on the language used, but also on contextual factors such as intent, audience, and impact. Given the lack of a specific legal provision on hate speech in Germany, research on classifying potentially criminal hate comments must account for this legal fragmentation. A suitable dataset must be aligned with judicial criteria rather than broad definitions from international organizations or private institutions. This underscores the importance of interdisciplinary approaches that combine computational linguistics, legal analysis, and social sciences to develop reliable automated models for detecting hate speech. For the purposes of this study, we define hate speech as verbal or written communication that denigrates, insults or threatens individuals or groups on the basis of characteristics such as ethnicity, nationality, religion, sexual orientation or political affiliation, which potentially constitutes an offence under one or more sections of the German Criminal Code (StGB). This definition emphasises legal applicability and is derived from previous research into legally relevant hate speech (Schäfer, 2023; Zufall et al., 2019)

2.4 Related Methods

In addition to the growing number of annotated datasets, various methodological approaches have been proposed for the detection of hate speech. Early work relied on lexicon-based methods and traditional classifiers, such as support vector machines (SVMs) or decision trees (Schmidt and Wiegand, 2017). However, with the advent of deep learning, recurrent neural networks, and, more recently, transformer-based models such as BERT and RoBERTa, classification performance has significantly improved (Mozafari et al., 2019) These models enable the contextual understanding of hate speech, which is essential for addressing nuanced legal categories.

3 Creation of a Dataset Aligned with Legal Classification

3.1 Requirements and Sources

The comments are collected from various sources for two primary reasons: First, to minimize bias. A single dataset may contain inherent biases based on the type of speech and the community from which the comments originate. By aggregating data from diverse sources, we ensure a more balanced representation, regardless of the comment's origin. Second, to ensure a sufficient number of comments for training a robust classifier, as many datasets lack a sufficient quantity of illegal instances necessary for effective model training.

The data sources used are as follows:

1. DeTox is a large dataset of Twitter messages with about 10.000 comments annotated for sentiment, toxicity, hate speech, discrimination, and legal relevance (Demus et al., 2022). Based on this annotation, we identified 385 comments that are likely to be illegal, with a match rate of at least 0.67. We also identified 300 hate comments that were randomly selected, with a match rate of 1.0, and that had no apparent criminal relevance.

2. IHS is another dataset of Twitter messages containing potentially illegal hate speech and annotated according to the applied criminal law sections/groups of sections (Schäfer, 2023). The data were annotated by a single trained person. 287 comments with an assigned criminal relevance were retrieved.

3. The X platform is an established source for data collection. An exploratory analysis was conducted using the platform's search function, with the keywords listed in Table 5 being utilised to obtain an up-to-date overview of criminally relevant content. In the course of this preliminary investigation, 93 public comments were identified and included in the data set that potentially fall under the criminal provisions of Sections 86 and 86a of the German Criminal Code (StGB).

4. 125 comments generated by the GPT-3.5 model constituted the final part of the dataset. The model was given seed examples derived from real-world hate speech comments (see Table 5) and instructed to produce similar utterances. The primary motivation for including synthetic data was to augment the existing dataset in a controlled and targeted manner. This approach aimed to enrich the dataset with additional, diverse examples of hate speech. Data augmentation through large language models has proven effective in various NLP tasks, including toxic language classification, as it allows for scalable generation of realistic yet varied training samples (Jahan et al., 2024).

Note that during the data collection phase, we focused on retrieving comments labeled as illegal in their original annotations. However, our final annotations differ significantly from the original dataset annotations and also from our initial assessments. This discrepancy may be attributed to the lack of quality in the retrieved datasets, as the original annotation process did not involve legal professionals or experts and was conducted by students from non-legal fields (Schäfer, 2023; Demus et al., 2022). Another potential reason for the mismatch in annotations could be the inherent complexity of the annotation task, as we will discuss in Section 3.4.

In total, the complete data set has 1,090 comments. The German-language examples have been adopted in their original spelling without correction; the respective English translations are given in the footnotes. The final version of the data set is open source, with hidden usernames (replaced with 'user'). No further modifications were made to the comments.

3.2 Dataset Structure

We began by collecting sections from StGB that are likely to be applied to the written illegal comments. Those sections were grouped into three classes, according to the offense they constitute, under the supervision of public prosecutors. Such a classification is essential givem the strong similarities in the sections' content within one group, which makes it challenging to distinguish between them without professional expertise. For instance, the following comment could be either interpreted as insult (§ 185 StGB) or malicious gossip (§ 186 StGB). The term 'corrupt' suggests malicious gossip, while 'puppet pig' constitutes an insult:

> 'Scholz ist für uns Sachsen nicht existent ! Wir haben dieses korrupte, hochkriminelle, kommunistische Marionettenschwein nicht gewählt! Wie wollen nichts mehr mit Berlin und Brüssel zu tun haben!'[1].

By organizing the sections into these classes, we ensure high annotation quality. The fourth class we add includes hate comments that presumably do not contain criminal offenses of the target sections. The full list of classes and the corresponding sections is as follows:

1. Class 1: Dissemination of unconstitutional Material

   (a) §86 StGB Distributing propaganda materials from unconstitutional organizations

   (b) §86a StGB Use of symbols of unconstitutional organizations

   The comments of this class disseminate propaganda material or symbols (such as slogans or forms of greetings) of banned organizations or unconstitutional political parties. For example, the comment

   > 'Wir müssen unsere Führer unterstützen und unsere Opposition bekämpfen. #BlutUndEhre'[2]

   contains, among other elements, the phrase 'Blut und Ehre'[3], which was the central motif of the German Hitler Youth and refers to a neo-Nazi network that was banned in Germany in the year 2000.

2. Class 2: Public Incitement to Commit Crimes and Disturbing the Public Peace

   (a) §111 StGB Public incitement to commit crimes

   (b) §126 StGB Disturbance of public peace through the threat of criminal offenses

   (c) §130 StGB Incitement of masses

   (d) §131 StGB Depictions of violence

   (e) §140 StGB Rewarding and approval of offenses

   This class includes comments that disturb public peace. More specifically, they incite hatred, violence, or criminal acts; attack the dignity of a large group; glorify or downplay inhuman violence or Nazi crimes. It also covers threats of serious offenses and public approval of severe recent crimes. An example of a comment that clearly falls into this category is:

   > 'Diese Kriminellen in Der Medizin gehören mit Genickschuss hingerichtet. #Nürnberger Kodex'[4],

   which explicitly incites murder in connection with an implicit reference to the methods of the Nazi regime.

[1] Scholz is non-existent for us Saxons! We did not elect this corrupt, highly criminal, communist puppet pig! We want no further involvement with Berlin and Brussels!

[2] We must support our leaders and fight our opposition. #BloodAndHonour

[3] Blood and Honour

[4] These criminals in medicine should be shot in the neck. #Nuremberg Code

3. Class 3: Defamation and Insult
   (a) §166 StGB Revilement of religious faiths and religious and ideological communitiess
   (b) §185 StGB Insult
   (c) § 186 StGB Malicious gossip

   The comments of this class insult or degrade individuals or small groups, including religious communities. For instance, the following comment contains a strong personal insult against three users:

   > ‘@user @user @user Fresse halten, asoziales dummes Stück Vieh.’[5]

4. Class 0: Legal/Does not Belong to the Aforementioned Classes

   > ‘Denn was Hatespeech ist, bestimmen irgendwelche grün-links*extremen S. p. i. .n. N. e. r.....’[6]

This comment falls into class 0, as the phrase does not explicitly target a specific individual or legally protected group in a manner that would meet the criteria for criminal relevance.

3.3 Dataset Annotation

The data set was annotated in two phases. In the first phase, six public prosecutors from the Central Office for Combating Internet and Computer Crime (ZIT), who handle reported hate comments daily, annotated 351 comments. The second phase involved an intensive workshop with the participation of public prosecutors and employees of the hate speech reporting center "Hesse Against Hate Speech" who provided valuable insights into the comment annotation process, serving as the foundation for a comprehensive annotation guideline. Drawing on insights from the workshop and cross-referencing the experts’ annotations, two groups — each consisting of three native German students and academic staff — annotated an additional 839 comments. The Fleiss’ Kappa score before the collective discussion was 0.608 and 0.811 for the two groups. Each group resolved disagreements through discussion until a consensus was reached. According to the annotation rules, the comment could be assigned to one or several classes, depending on the range of possible offenses it carried. However, if the comment belonged to Class 0 (i.e., was deemed legal), it could not be assigned to any other class. The final annotation statistics are given along with their total number per class in Table 3. To provide a clearer picture, Table 2 presents several examples from the dataset along with their offenses and the corresponding classes.The multilabel annotation proved to be important, with 7 percent (88) of the comments belonging to multiple classes, occasionally, to all three of them (see Table 1).

Classes	Amount
Class 1&2	24
Class 1&3	6
Class 2&3	55
Class 1&2&3	3

Table 1: Comments Belonging to Multiple Classes

3.4 Challenges in Dataset Annotation

The line between comments that are covered by freedom of speech and those that are legally relevant is not always clearly defined, and therefore is often subjective. In this context, we compare example sentences that at first look very similar but have been evaluated differently by experts in terms of their criminal relevance.

1. Use of a banned organization symbol:
   (a) H**l Hi**er! Wir sind die Zukunft Deutschlands![7] – *Class 1*
   (b) Rotfront ver**cke, ich zeige ihnen, was ich bin[8] – *Class 0*

   ‘Rotfront ver**cke’ is an exclamation from the German National Socialist era, directed against the German Communist Party. Despite this, the isolated use of the expression is not necessarily to be regarded as illegal, since there is no explicit right-wing extremist context, the organization has ceased to exist and the expression can only be subjectively associated with an anti-socialist attitude. Without investigating the author’s profile and history, no definitive legal judgment can be made. In turn, ‘H**l Hi**r’, a slogan of the banned

[5] @user @user @user shut up, antisocial stupid piece of cattle.

[6] Because what constitutes hate speech is determined by some green-left*extremist w. e. i. r. d. o. s.....

[7] H**l Hi**r! We are the future of Germany!

[8] Rotfront ver**cke, I’ll show them what I am

Comment	Violations of Law	Class
@user @user Welch eine hirnlose dumme Antwort von Ihnen!!![13]	Does not violate the law	0
H**l Hi**er, mein Herz bleibt immer für Deutschland[14]	Disseminates the symbol of a former National Socialist organization	1
@user gehört aufgehängt[15]	Incites violence and a criminal act	2
@user darf ich dich h**e nennen[16]	Insults an individual	3
@user:deine Zeit kommt noch: wir kriegen dich schon. H**l Hi**er![17]	Threatens to commit a dangerous bodily harm, disseminates the symbol of a former National Socialist organization	1,2
heil hitler du nut**nsohn[18]	Disseminates the symbol of a former National Socialist organization, insults an individual	1,3
Eine Kugel reicht für dieses matschhirn[19]	Threatens to commit a dangerous bodily harm, insults an individual	2,3
Liebe @user, [u]nser Freund *Name* hat [...] zu seiner Waffe gegriffen, um Untermenschen auszulöschen. [...] Wir werden euch auch noch dran kriegen und euch Genicksch**se schenken. Ihr seid unsere Feinde [...] Die nächsten Anschläge sind bereits in Planung [...] S**g H**l und H**l Hi**er! [20]	Publicly approves a commited crime under aggravating circumstances, threatens to commit a criminal offense constituting a public danger, incites to hatred and calls for violence against national, racial, and religious groups, disseminates the symbol of a former National Socialist organization, commits an insult	1,2,3

Table 2: Examples, violations and corresponding classes created by the authors in the framework of the research.

party NSDAP, is treated more severely and always constitutes an offense.

2. Incitement to hatred

   (a) die MOSLEMS gehören alle mit einer hitlersäge zersch**sen! habe die ehre.[9] – *Class 2*

   (b) Ausländer/Türken/Muslime raus![10] – *Class 0*

   Although calls for deportation are often associated with far-right hate groups, the isolated use of such rhetoric is not generally considered to be illegal and is protected under the right to freedom of expression. However, severe insults inciting hatred, calls for violence against the aforementioned groups, and attempts to disrupt public peace are subject to legal prosecution.

3. Insult

   (a) @user du ekelhaftes fettes Schwein[11] – *Class 3*

   (b) sein Ego nur durch Hass, Rassismus, Fremdenfeindlichkeit und Hetze aufwerten kann, ist ein ganz winziger Wicht und feiger Hosensch**ßer. Du bist halt 'ne lappige Fritte.[12] – *Class 0*

   The distinction between expressing an opinion and insulting someone's personal dignity is often vague. However, criticism of a person's ideas and behavior is generally afforded greater protection than attacks on their physical appearance.

The comparison of comments illustrates that accurately detecting offenses requires extensive ex-

[9] The Muslims should all be sh*t with a Hitler saw! Have the honor.

[10] Foreigners/Turks/Muslims out!

[11] @user you disgusting fat pig

[12] [someone who] can only inflate their ego through hate, racism, xenophobia, and incitement is a tiny little nobody and a cowardly pants-sh**ter. You are just a limp fry

[13] @user @user What a mindless stupid answer from you!!!

[14] H**l Hi**er, my heart will always be with Germany

[15] @user should be hung up

[16] @user may I call you a wh**re

[17] @user your time will come: we'll get you. H**l Hi**er!

[18] h**l hi**er you son of a b**ch

[19] One bullet is enough for this mushy brain

[20] Dear @user, [O]ur friend *Name* [...] grabbed his weapon to exterminate subhumans. [...] We will get you too and give you ex**ution shots. You are our enemies [...] The next attacks are already being planned [...] S**g H**l and H**l Hi**er!

Annotator Group	Class 1	Class 2	Class 3	Class 0	**Total per Group**
ZIT Group	47	68	89	158	351
Trained Group 1	16	121	119	214	416
Trained Group 2	125	26	13	276	423
Total per Class	188	215	221	648	

Table 3: Annotation Statistics

perience in this field. While all six experts were unanimous in their annotations, the student group had disagreements even after discussion (these examples were excluded from the final version of the dataset). For example the following comment:

> 'Verdient, wenn die Deutsche Polizei sich weiter so rassistisch verhält kein Wunder wenn dann Polizisten ermordet werden. Selbst Schuld Deutsche Polizei, ihr schießt euch ins eigene Kopf. Der Täter verdient von mir Respekt, unendlich Respekt, küsst dem *NAME* die Füße, obwohl bei Deutschen Tätern nennt ihr es Physiche Probleme aber bei Ausländern ? Was sagt ihr Rechtsextremen, dass es dann ein Terrorist ist ? Schämt euch, ich hoffe mehr solcher Fälle passieren bitte'[21]

On one hand, the comment might be assigned to Class 2, 'Public Incitement to Commit Crimes and Disturbing the Public Peace', as it contains calls for violence against police officers and, moreover, rewards the crime committed. On the other hand, to confidently annotate the comment as illegal, additional surrounding context may be needed. For example, it is unclear whether the comment rewards a recently committed crime or one from the past, whether the author has made other hateful statements that could reveal their intentions, or whether they belong to any extremist groups. Ultimately, no agreement was reached between the annotators.

4 Experiments

To demonstrate the usability of the newly annotated dataset, we introduce a baseline classifier as an initial benchmark. This baseline ensures that the dataset provides sufficient signal for meaningful learning and serves as a reference for future models. Specifically, we assess the dataset using an in-context learning approach with the open foundation LLM LLaMA-70B, using the hyperparameters specified in Table 6. In in-context learning, relevant training examples are incorporated directly into the prompt to guide the model's predictions. We conducted experiments using few-shot learning paradigms (Liu et al., 2021). The training examples were selected using a k-nearest neighbor (KNN) approach to identify semantically similar comments. To obtain embeddings for measuring similarity, we fine-tuned a RoBERTa model on the entire training corpus using the hyperparameters specified in Table 7 (shown in Appendix). Furthermore, we evaluated two prompt variations: one including class descriptions (Table 8) and another without them (Table 9). We hypothesize that providing explicit class descriptions will equip the LLM with sufficient context to enhance its predictive accuracy while ensuring awareness of all possible classes. In contrast, the prompt variation without class descriptions requires the model to infer the number and nature of classes implicitly from the context of the given demonstrations. To eliminate dependence on a specific dataset split, we employed 10-fold cross-validation on the entire dataset in all experiments. For each fold, the training data was used to fine-tune a RoBERta model, while the corresponding held-out fold served as the evaluation set.

Given the legal complexity and multi-label nature of our task, prompt design is particularly important. Including class descriptions in the prompt enables the model to distinguish between overlapping legal categories more effectively, thereby reducing ambiguity and improving classification precision. This approach aligns with findings that

[21] Deserved, if the German police continue behaving so racially, no wonder when police officers are then murdered. The German police are to blame, you're shooting yourselves in the head. The perpetrator deserves my respect, infinite respect, kisses *NAME*'s feet, even though with German perpetrators you call it psychological problems, but with foreigners? What do you right-wing extremists say, that they're terrorists? Shame on you, I hope more of these cases happen, please

structured, informative prompts can elicit more accurate reasoning and output from LLMs. By comparing the model's performance with and without class descriptions, we provide empirical evidence of the importance of prompt clarity in multi-label classification scenarios involving nuanced legal distinctions.

5 Results

5.1 Baseline Classification Performance

Table 4 presents the classification results obtained from this procedure for the LLM with and without class descriptions. Interestingly, the results indicate that the classifier achieves a better overall performance when using an implicit prompt rather than an explicitly structured one with class descriptions. This finding contrasts with initial expectations, since structured prompts were assumed to provide clearer guidance for the model. Nevertheless, the classifier consistently distinguishes well between legal and illegal comments, particularly excelling in detecting legally irrelevant content (Class 0). However, performance slightly declines for more nuanced categories, such as defamation and insult (Class 3), suggesting that further refinements in feature representation and annotation strategies could enhance classification accuracy.

5.2 Effect of Prompt Engineering

We further analyzed the impact of different prompt variations on the performance of the model. Two variations were tested: (1) a structured prompt explicitly defining all class descriptions and (2) an implicit prompt requiring the model to infer classes from provided examples. Contrary to expectations, explicit prompt formulation resulted in lower accuracy, highlighting the challenges of providing structured class descriptions for legal text classification tasks. This finding suggests that future models might benefit from a more implicit approach to class inference, allowing greater adaptability and context awareness, as clearly seen in Figure 1.

5.3 Multilabel Classification Analysis

A key aspect of our dataset is the ability to assign multiple legal classifications to a single comment. We observed that 7% of the comments fell into multiple categories, some even exhibiting all three legally relevant offenses simultaneously. The breakdown of multi-label cases is presented in Table 1. These results highlight the complexity of legal text classification, where multiple legal violations often coexist within a single statement. This underscores the need for robust annotations that capture overlapping legal offenses.

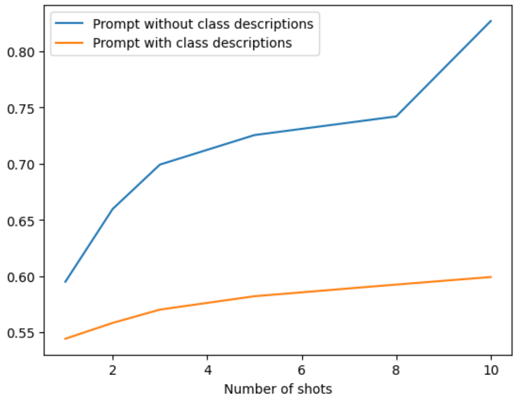

Figure 1: Accuracy over number of shots

In-context learning (ICL) has emerged as a powerful paradigm for utilising large language models without requiring parameter updates (see, for example Brown et al. (2020) and Min et al. (2022)). By conditioning the model on a sequence of input–output examples, known as demonstrations, ICL enables flexible adaptation to new tasks and domains. Recent research has demonstrated that the selection and phrasing of prompts have a significant impact on model performance, a field commonly referred to as prompt engineering (Liu et al., 2021). Our prompt design is therefore informed by insights from the literature on prompt engineering, particularly with regard to class disambiguation and interpretability (Wei et al., 2021).

Given the legal complexity and multi-label nature of our task, prompt design is particularly important. Including class descriptions in the prompt enables the model to distinguish between overlapping legal categories more effectively, thereby reducing ambiguity and improving classification precision. This approach aligns with findings that structured, informative prompts can elicit more accurate reasoning and output from LLMs. By comparing the model's performance with and without class descriptions, we provide empirical evidence of the importance of prompt clarity in multi-label classification scenarios involving nuanced legal distinctions.

6 Conclusions

In this study, we introduced a novel dataset for classifying legally relevant hate speech and conducted

	With class descriptions			Without class descriptions		
Nr Shots	Accuracy	Micro F_1	Macro F_1	Accuracy	Micro F_1	Macro F_1
1	0.54426	0.61877	0.64894	0.59496	0.64754	0.65023
2	0.54426	0.62973	0.65967	0.65966	0.71317	0.70263
3	0.57017	0.63968	0.66924	0.69916	0.75263	0.73491
5	0.58207	0.64883	0.67708	0.72542	0.77834	0.75693
8	0.59244	0.65833	0.68361	0.74202	0.79486	0.77155
10	0.59916	0.66295	0.68732	0.82689	0.87452	0.85230

Table 4: Performance of the classifier with explicit class descriptions at different shot counts.

extensive experiments to evaluate classification performance. Our results highlight the effectiveness of in-context learning and demonstrate the importance of careful annotation and prompt engineering. While our baseline model performed well in distinguishing legal from illegal content, challenges remain in accurately classifying nuanced legal offenses. Key findings suggest that implicit prompt-based classification methods may yield better accuracy than explicitly structured prompts, emphasizing the need for further research into context-aware classification models. Additionally, our multilabel classification approach underscores the complexity of legal text classification, where multiple offenses often overlap. Despite promising results, limitations such as annotation subjectivity, class imbalance, and interpretative challenges must be addressed in future work. Enhancing dataset diversity, refining annotation protocols, and incorporating expert-driven methodologies will further improve classification robustness. Ultimately, this research represents progress toward the development of automated legal classification tools that can assist law enforcement agencies and legal practitioners in identifying criminally relevant online discourse. Future efforts should focus on improving model interpretability, reducing biases, and ensuring ethical considerations in the deployment of AI-driven legal assessment systems

Limitations

Despite the involvement of legal experts, the classification of hate comments into legal categories remains a complex task with occasional ambiguity, even among professionals. Additionally, the dataset exhibits a certain imbalance, with some categories, such as Class 1 and Class 3, containing significantly fewer examples, which could potentially affect model training. Furthermore, while a range of prompts and German language models were considered, the limited number of prompt formulations and models used in data generation has led to a degree of homogeneity in the AI-generated hate speech data. To address potential model-specific biases, the test data includes human-authored comments as well as data from Mixtral-8x7B, a model excluded from the training set. Lastly, the annotations of the dataset are tailored to German law, which may limit its applicability to the legal systems of other nations.

References

Markus Bertram, Johannes Schäfer, and Thomas Mandl. 2023. Comparative survey of German hate speech datasets: Background, characteristics and biases. In *Lernen, Wissen, Daten, Analysen*.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Hawaii International Conference on System Sciences*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bundeskriminalamt. 2024. Zentrale Meldestelle Hasskommentare BKA. Daten und Zahlen.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don‘t be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

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

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.

Digital Service Act. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 october 2022 on a single market for Digital services and amending directive 2000/31/ec (Digital Services Act) (text with eea relevance): Digital Services Act. Accessed on L 277/4.

European Commission against Racism and Intolerance (ECRI). 2024. ECRI general policy recommendation no. 15 on combating hate speech.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. *Preprint*, arXiv:2403.19559.

HessenGegenHetze. 2025. Fünf Jahre Meldestelle Hessengegenhetze.

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

Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddia, Jhuma kabir Mim, and Nabil Arhab. 2024. A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Meta. 2024. Hate speech.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media.

Björn 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. *ArXiv*, abs/1701.08118.

Johannes Schäfer. 2023. Bias mitigation for capturing potentially illegal hate speech. *Datenbank-Spektrum*, 23:41–51.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Johannes Schäfer. 2023. Bias mitigation for capturing potentially illegal hate speech. *Datenbank-Spektrum*, 23.

StGB Strafgesetzbuch. 2024. § 107c StGB.

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.

United Nations (UN). 2024. What is hate speech?

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Le Quoc, and Denny Zhou. 2021. Chain-of-thought prompting elicits reasoning in large language models.

Michael Wiegand and Melanie Siegel. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop*.

Frederike Zufall, Tobias Horsmann, and Torsten Zesch. 2019. From legal to technical concept: Towards an automated classification of German political Twitter postings as criminal offenses. In *North American Chapter of the Association for Computational Linguistics*.

Illegal Phrase	Illegal Nature
Kampf ist Teil des Lebens; er ist naturnotwendig für alles Werden, Sein und Vergehen. Jeder einzelne von uns wie unsere gesamte Art stehen in diesem Ringen.Wir bekennen uns zu diesem nie endenden Lebenskampf.	Doctrine of Artgemeinschaft
Weiße Revolution ist die einzige Lösung #318	Slogan of Combat 18
scheiß judenrepublik	Slogan of Wolfsbrigade 44
meine Ehre heißt Treue	Slogan of Nazis
Statt Migration sollte das oberste Ziel der Regierung die Sicherung der Existenz unseres Volkes und die Zukunft für weiße Kindern sein 88	14 Words, Slogan of Neonazis
Alles für Deutschland	Slogan of Nazis
HFFH	Short for Hammerskins Forever, Forever Hammerskins
Heil Hitler!	Nazi Salute
Hammerskins forever, forever Hammerskins	Hammerskins Slogan
Für den Führer 1488	Combination of 14 Words and a Short Form of Heil Hitler
Rassenkampf für das weiße Volk C18	Reference to Combat 18
Ein Volk, ein Reich, ein Führer!	Slogan of Nazis
Stoppt die Islamisierung, Deutschland erwache	Slogan of Nazis
Unsere Ehre heißt Treue	Slogan of Nazis
Blut und Ehre	Slogan of Nazis
Rotfront verrecke	Slogan of Nazis
juden verreckt. jedem das seine	Was Written at the Entrance Gate to Buchenwald
Deutschland ist erwacht 1488 ... mit Blut und Ehre	Slogan of Nazis
Jeder Jude ist ein Siedler, und es ist unsere Pflicht, ihn zu töten.	Hamas Slogan
Israel existiert und wird weiter existieren, bis der Islam es ausgelöscht hat, so wie er schon andere Länder vorher ausgelöscht hat	Hamas Preamble

Table 5: Key Phrases for Twitter Search

Model Name	Llama 70b Instruct
Temperature	0.1
Top_p	1
Max Tokens	512

Table 6: Llama 70B hyperparameters used during inference

Parameter	**Value**
Number of Epochs	5
Batch Size	16
Learning Rate	4×10^{-5}
Weight Decay	0.05

Table 7: Fine-tuning hyperparameters for RoBERTa

You are an assistant, and your task is to evaluate the comments below based on the paragraphs of StGB.
Provide only the class number. Here are the classes:
Class 0: Not legally relevant
Class 1: Dissemination of unconstitutional material
Class 2: Public incitement to commit crimes and disturbance of public order
Class 3: Defamation and insult
If a comment is legal, it only belongs to Class 0. If a comment falls under multiple legally relevant classes, list them all.
The format of your response is: 'Class {Number}' or 'Class {Number_1}, Class {Number_2}'.

Table 8: Prompt with Class Descriptions

Task: Assign a class or classes to the comment based on examples.
Format of your response:

- *For one class: Class {Number}*
- *For multiple classes: Class {Number_1}, Class {Number_2}*

Table 9: Prompt without Class Descriptions