

FI-CODE@GermEval Shared Task 2025: LLM Prompting for Augmentation of Underrepresented Classes

Nina Seemann, Yeong Su Lee, Hendrik Bothe, Michaela Geierhos

University of the Bundeswehr Munich, Research Institute CODE

Werner-Heisenberg-Weg 39, 85579 Neubiberg, Germany

{nina.seemann, yeongsu.lee, hendrik.bothe, michaela.geierhos}@unibw.de

Abstract

This paper summarizes our participation in the Harmful Content Detection in Social Media shared task. The task is divided into three subtasks: Call2Action (C2A), Attacks on the Democratic Basic Order (DBO), and Violence Detection (VIO). Following the organizers' recommendation, we used data augmentation to create more instances of each underrepresented category. To this end, we asked three large language models to rewrite a tweet in six different styles. We used a German BERT model for classification. The macro F1 scores obtained for the three subtasks were 79 % for C2A, 62 % for DBO, and 75 % for VIO.

1 Introduction

Social media platforms have become essential spaces for political communication and public discourse (Stieglitz and Dang-Xuan, 2013; Sapar et al., 2023), as evidenced by the frequent posting of many politicians who use platforms like Twitter (now X) and Instagram to communicate daily with voters and influence media cycles. However, these platforms also facilitate the rapid dissemination of hate speech. Political actors, interest groups, and polarized communities may exploit these platforms to spread harmful, discriminatory, or extremist content – often under the guise of opinion or satire. The virality and reach of such messages can undermine democratic dialogue, fuel polarization, and marginalize vulnerable groups. The volume and speed of online content makes it difficult for manual moderation to keep pace, so there is a growing demand for automated hate speech detection systems. These tools are crucial for safeguarding users, fostering respectful discourse, and upholding the integrity of democratic processes, particularly during election campaigns and times of heightened political tension. However, developing robust models remains challenging due to the implicit, context-

dependent, and culturally varied nature of online hate speech.

In this Harmful Content Detection shared task, researchers focus on modeling three subtasks of such applications (Felser et al., 2025). Shared tasks on harmful content detection typically provide datasets containing normal and harmful content to differentiate between these two categories. These datasets primarily contain content from far-right extremist groups, as well as comments from non-supporters criticizing that content. Furthermore, not every tweet by a follower of an extremist group is automatically extremist. Hence, the goal is to identify content of the following three subtasks. Call2Action (C2A) aims to automatically identify calls to action on social media. This subtask involves a coarse-grained, binary classification. Attacks on the Democratic Basic Order (DBO) aims to automatically detect whether a tweet constitutes an attack on the free democratic basic order and involves classifying tweets into one of four categories. Violence Detection (VIO) aims to automatically determine if a tweet expresses an overly positive attitude toward violence. This subtask involves a coarse-grained, binary classification. For more details on each subtask, see Felser et al. (2025).

Harmful content on the internet is typically less numerous than normal content. This fact is reflected in the datasets provided for this shared task. Therefore, the organizers explicitly encouraged participants to use data augmentation, which we achieved by leveraging large language models (LLMs). Using LLMs for data augmentation is not a new idea (Ding et al., 2024; Jahan et al., 2024). However, in the context of this shared task, this approach may be ineffective due to the offensive and disturbing nature of the tweets. LLMs are trained on online texts containing harmful content¹, so cre-

¹<https://research.ibm.com/blog/efficient-llm-hap-detector> (2025-07-15)

ators are concerned about teaching their LLMs not to replicate or filter this content. Nevertheless, to obtain more instances for each underrepresented class, we asked three LLMs to rewrite a given tweet in six different styles.

In our experiments, we focus on two things. First, we examine the ability of LLMs to rewrite disturbing texts. Second, we examine how synthetic data can be helpful. Our goal is not to obtain the best F1 scores. This paper is organized as follows: Section 2 provides an overview of related work. In Section 3, we present the overview of the datasets provided by the organizers. Section 4 shows the process of data augmentation, while Section 5 details our experimental setups and results, and discusses related issues. Finally, we conclude our research and present future work in Section 6.

2 Related Work

Data augmentation is an established technique for improving the performance of machine learning models when working with limited or imbalanced datasets. Traditional methods, such as SMOTE, generate synthetic samples by interpolating between existing data points in the feature space (Chawla et al., 2002). However, they often fail to capture the semantic complexity of textual data. More recent approaches use language-based techniques, such as synonym replacement, back-translation, and contextual embeddings, to introduce linguistic variation (Wei and Zou, 2019; Ciolino et al., 2022; Pantelidou et al., 2022; Jahan et al., 2024). Thanks to powerful LLMs, a new approach has emerged that generates semantically coherent augmented data.

Studies have shown that text generated by LLMs can significantly improve classification performance in scenarios involving short texts (Balkus and Yan, 2023). Compared to traditional augmentation methods, LLMs offer more linguistic diversity and are better at capturing domain-specific nuances (Ding et al., 2024). This is particularly important for classifying posts on platforms such as Twitter (now X), where informal language, irony, and brevity present additional challenges. The same applies to hate speech detection. The effectiveness of these models is closely related to the quality and diversity of their training data. However, recent studies have revealed that commonly used hate speech datasets exhibit systematic issues in their composition, such as imbalanced label dis-

tributions, platform-specific sampling biases, and inconsistent annotations (Jaf and Barakat, 2024). Models based on these biased datasets can learn biased patterns and produce unreliable results.

Therefore, recent research addresses these challenges by focusing on context-aware data augmentation strategies aimed at improving the representation of minority classes. The GAHD dataset, a new collection of approximately 11,000 instances of adversarial hate speech in German, was introduced to evaluate model robustness (Goldzycher et al., 2024). Experiments demonstrate that GAHD poses a significant challenge, even to state-of-the-art detection models. Incorporating GAHD into the training process notably enhances classification stability and generalization.

Other researchers have used different augmentation approaches to improve hate speech detection performance. A combination of synonym and token replacement, paraphrasing, and back-translation has been shown to significantly improve model performance when used as data augmentation techniques (Aljawazeri and Jasim, 2025). Jahan et al. (2024) evaluated several data augmentation techniques for hate speech detection, with a particular focus on LLM generation. Among these techniques, GPT-3-based augmentation was shown to generate syntactically rich and semantically coherent samples, thereby improving model performance. Unlike traditional methods, such as back-translation or synonym substitution, LLM-based augmentation enriches the diversity of the training distribution. The findings suggest that incorporating augmented data substantially improves classification accuracy and generalization, especially in low-resource scenarios.

Others explored the automated classification of Calls to Action in Instagram content from the 2021 German federal election campaign (Achmann-Denkler et al., 2024). They combined fine-tuned BERT (Devlin et al., 2019) models with synthetic training data generated via LLMs, specifically GPT-4. Using augmented data enabled them to achieve high classification accuracy and revealed distinct mobilization strategies across content types (posts vs. stories) and political parties. These results highlight the importance of LLM-driven augmentation in improving the robustness and generalization of real-world social media classification tasks.

3 Datasets

The organizers provided a newly annotated dataset for the aforementioned tasks (Felser et al., 2025). Tweets were collected between 2014 and 2016 and annotated by students or research assistants who are experts in identifying the characteristics of harmful textual content. The dataset consists of training and trial data. The trial data itself is included in the training data, hence we removed the trial part from the training data for our experiments (see Section 5). Table 1 gives a detailed overview of the dataset sizes per label for each subtask after excluding trial from training data.

As the table shows, there is significant imbalance in label distribution. Therefore, we augment each underrepresented class with LLMs.

Subtask	Label	# Instances	
		Train	Trial
C2A	true	561	102
	false	5,228	949
DBO	subversive	56	4
	agitation	307	6
	criticism	683	122
	nothing	5,653	921
VIO	true	504	60
	false	6,228	991

Table 1: An overview of the sizes of the training and trial data for each subtask.

4 Data Augmentation

Below, we describe the LLMs we used for data augmentation and the rewriting styles we selected. The augmented data was generated on a single Nvidia A100 80GB GPU, with a processing time of approximately 5 seconds per tweet.

4.1 Large Language Models

We utilized the following three state-of-the-art language models: Llama 3.3 with 70 billion parameters (Grattafiori et al., 2024), Gemma 3 with 27 billion parameters (Team et al., 2025) and Phi-4 with 14 billion parameters (Abdin et al., 2024) for data augmentation, as they represent diverse architectural paradigms and parameter scales, enabling the generation of robust, varied, and semantically coherent synthetic data.

The rewritten tweets are generated via the Ollama Python library², which provides a common interface for many published LLMs. A new model is initialized for each LLM and writing style using the same Ollama procedure, where only the LLM and writing style names are changed. First, we specify the LLM and the task that the model should perform. This task instructs the model to interpret all subsequent inputs as tweets to be rewritten in the specified tone. In addition to the tone, the LLMs were instructed to exchange entities. For example, the entity ‘floor’ could be exchanged for ‘carpet’. This exchange should further diversify the dataset. LLMs typically refuse to answer questions about specific tasks involving harmful speech. Therefore, the instructions include a sentence stating that the task is for research purposes only, to minimize refusals to answer. If an LLM cannot interpret a tweet, it was instructed to return the word ‘None’. The chosen parameters, including the prompt, can be found in Appendix A. After initialization, each tweet from an underrepresented class is fed into the LLM. These responses represent the augmented data. Depending on the LLM, some tweets were rejected or misinterpreted, rendering the responses unusable as augmented data. The cleaning of the created datasets is discussed in Section 4.3.

4.2 Rewriting Styles

The styles of social media posts have been widely studied. On Twitter (now X), for example, styles range from formal to casual and intimate (Sani et al., 2025). Their analysis revealed that the casual style was the most prevalent among Twitter users, accounting for approximately 40 % of the sampled tweets. This style characterizes the informal and conversational nature of the Twitter platform. Hu (2023) analyzed political tweets and identified six language styles: mobilizing, self-expressive, argumentative, narrative, analytical, and informative. The study examined how these styles relate to users’ online characteristics, such as social connections or gender. Thus, we asked LLMs to identify the speech styles from a range of predefined styles in the tweets of the provided dataset. We selected these styles without conducting any theoretical analysis. Table 2 shows the prompt with the parameters.

The overall results are shown in Table 3, separated by class. However, it is worth noting the

²<https://www.ollama.com> (2025-07-15)

request	
role:	system
content:	Your task is to guess the speech style for the input text. Give only the guessed speech style without any comments or explanation. You have the following choices. Give one of these choices as answer back: {Neutral, Formal, Polemical, Sarcastic, Satirical, Aggressive, Polite, Colloquial, Ironic, Poetic, Biting, None}. If you do not have any text to analyze the style, give just None as answer back!
parameter:	{temperature: 0.9, top_p: 0.9}
role:	user
content:	Ist das nicht die Bürgermeisterin, die gegen den Drogenhandel im Görlitzer Park die Büsche abholzen liess? Durch solche Leute können diese Zustände erst entstehen.
response	
role:	Assistant
content:	Polemical

Table 2: Example of a prompt request and the response returned for a style guess.

difference in distribution between the negative and positive classes for all tasks. The colloquial style is most prevalent in negative tweets, while the aggressive style clearly dominates in positive tweets. We chose six of the returned guessed styles for data augmentation for the underrepresented classes: *bissig*, *formell*, *ironisch*, *neutral*, *polemisch*, and *sarkastisch* (The English translations are: *biting*, *formal*, *ironic*, *neutral*, *polemic*, and *sarcastic*). The following examples show the rewritten texts from Llama3.3:70b for the example text in Table 2.

- (1) Die selbsternannte Heldin des Görlitzer Parks, welche glaubt, dass das Abholzen von Büschen den Drogenhandel bekämpft, trägt mit ihrem unzureichenden Handeln direkt zu solchen katastrophalen Zuständen bei! [*biting*]
- (2) Die Entscheidung der zuständigen Kommunalbehörde, die Vegetation in einem bestimmten öffentlichen Park zu entfernen, um den Drogenhandel zu bekämpfen, wurde von einigen als maßgeblicher Faktor für das Entstehen bestimmter sozialer Probleme angesehen. [*formal*]
- (3) Die Heldin des Görlitzer Parks, die mit ihrem genialen Plan, die Büsche abzuholzen, den Drogenhandel sicherlich ausgerottet hat - ein perfektes Beispiel dafür, wie man Probleme wirklich löst. [*ironic*]
- (4) Eine lokale Politikerin hat Maßnahmen ergriffen, um den Drogenhandel in einem bes-

timtten Park zu bekämpfen. Kritiker sind der Meinung, dass solche Aktionen dazu beitragen könnten, Probleme zu verschärfen. [*neutral*]

- (5) Die Verantwortliche für die Rodung der Büsche im Görlitzer Park, um den Drogenhandel zu bekämpfen, trägt selbst zur Entstehung dieser abscheulichen Verhältnisse bei! [*polemic*]
- (6) Das ist bestimmt dieselbe Bürgermeisterin, die so clever war, die Büsche im Görlitzer Park abzuholzen, um den Drogenhandel zu bekämpfen - als ob das Problem einfach durch Entfernen von Pflanzen gelöst werden könnte. Es ist wirklich erstaunlich, wie solche Visionäre dazu beitragen, dass solche Zustände überhaupt entstehen können! [*sarcastic*]

In the dataset, the example in Table 2 is annotated as ‘criticism’. Rewriting such an example as ‘neutral’, might change the tone of the tweet and would be annotated as ‘nothing’. We ignore this fact and leave the original label as it is. We will discuss the effects at the end of Section 5.3.

4.3 Cleaning of Synthetic Data

As explained in Section 4.1 and demonstrated in the prompt in Appendix A, we instructed each LLM to output either the rewritten tweet or ‘None’ if rewriting was not possible. All answers containing ‘None’ were filtered out before training. Additionally, we manually inspected the answers to ensure the instruction was followed. During this process, we noticed that the LLMs handled our instruction very differently, as discussed further in this section.

Gemma 3. This LLM followed the instructions incredibly well. It either provided a rewritten tweet or answered ‘None’. For some tweets, it stated that they could not be rewritten due to their hateful content. These instances could easily have been filtered out beforehand. Additionally, there are only a few exceptions in which it began its response with “Here is the rewritten tweet” or something similar. In these cases, we removed that part of the answer and only added the rewritten tweet to the dataset.

Llama 3.3. Llama understood the instructions well and most of the time answered with either a rewritten tweet or ‘None’. There were some exceptions when it answered that it could not rewrite the tweet, providing an explanation for its decision. In these cases, the tweet either did not contain text (e.g., tweets containing only a URL) or

	C2A		DBO		VIO	
	False	True	Nothing	Others	False	True
Colloquial	2,179	126	2,285	90	2,417	90
Aggressive	1,247	317	1,542	490	1,507	353
Polemical	1,220	161	930	436	1,561	56
Sarcastic	958	34	818	135	1,009	54
Formal	82	8	116	5	114	4
Neutral	78	0	79	2	90	0
Polite	74	6	85	0	88	0
Satirical	67	4	52	10	67	6
Poetic	35	2	43	2	39	0
Ironic	19	0	20	3	21	0
Biting	4	1	2	1	3	0

Table 3: Results of the style guess per class for each subtask. Numbers in bold indicate the most prevalent style. Answers with ‘None’ are omitted.

the textual content was too disturbing, thus violating the LLM’s principles. All of these exceptions were filtered out before training. For some tweets, Llama provided a rewritten version but stated at the beginning of its answer that the tweet was of questionable character. We removed that part of the answer and added only the rewritten tweet to the augmented dataset.

Phi-4. A visual inspection of this LLM’s answers revealed that it completely ignored the instruction. One could describe it as “very chatty” with a lot of unnecessary output. Like the other two LLMs, it would sometimes answer that it could not rewrite a given tweet due to its hateful nature. However, even when it provided a rewritten tweet, Phi-4 also provided extensive additional information. For example, it provided an English translation of the tweet, the entities it replaced, the reasons for the replacements, a restatement of the task, and notes. Unfortunately, the additional output was not structured. Sometimes this output was marked with a star (*), and other times with a hashtag (#). Furthermore, the additional information was not always in the same position, i.e., at the beginning or end of the answer. During our inspection, the list of phrases to be filtered out became so long that we decided against using Phi-4’s augmented data.

Summary. Although LLMs are trained to avoid repeating harmful content, we were able to bypass this behavior by informing each LLM that the task was scientific in nature. Still, there were some instances in which they refused to rewrite a tweet,

though most of the time they complied. Gemma, in particular, rewrote more tweets than Llama. Table 4 illustrates this, showing the number of synthetic instances after cleaning. For DBO, we accumulated the numbers for each label. For a detailed overview of the synthetic instances per label for this subtask, see Table 10 in Appendix C.

5 Experiments & Results

In the following sections, we first describe the model that we used in our experiments. Then, we will describe the different setups that we evaluated. Lastly, we present the results for each subtask and setup.

5.1 Architecture & Model

For all experiments, we used the pre-trained GBERT base language model (Chan et al., 2020) from Huggingface³. The model was collaboratively trained by the creators of the original German BERT and dbmdz BERT models⁴. The following hyperparameters were applied: batch size = 8, learning rate = $2e^{-5}$, epochs = 3, and maximum sequence length = 128. Apart from the required preprocessing steps for BERT (adding [CLS] and [SEP] tokens, as well as tokenization with the BERT tokenizer), no other steps were implemented.

As mentioned in the introduction, our focus is on the ability of LLMs to rewrite harmful content and

³<https://huggingface.co/deepset/gbert-base> (2025-07-15)

⁴<https://github.com/dbmdz/berts>

Subtask	LLM	Style						Total
		<i>bissig</i>	<i>formell</i>	<i>ironisch</i>	<i>neutral</i>	<i>polemisch</i>	<i>sarkastisch</i>	
C2A	Gemma	657	661	658	660	657	659	3,952
	Llama	564	627	607	651	524	622	3,595
DBO	Gemma	1,167	1,169	1,174	1,170	867	1,174	6,721
	Llama	998	1,061	1,046	1,146	925	1,074	6,250
VIO	Gemma	555	558	561	559	560	561	3,354
	Llama	410	507	455	548	350	481	2,751

Table 4: Overview of the synthetic instances for each subtask, broken down by LLM and style.

the extent to which synthetic data is helpful in this shared task. Therefore, instead of training the 39 models on a GPU, we conducted all experiments on a MacBook Pro with an Apple M3 chip and 36 GB of RAM.

5.2 Experimental Setups

For each subtask, we first obtained baseline results using the provided training data (excluding the trial data as explained in Section 3). Then, we conducted preliminary experiments to determine the optimal combination of training data and LLM. We evaluated four different setups, which we will explain in greater detail. Section 5.3 provides details about the outcomes.

0 Baseline. To contextualize our results obtained by adding augmented data from LLMs, we first trained a baseline model. For training, we only used the provided datasets for each subtask.

1 Effectiveness of all styles from each LLM. In our initial setup, we evaluated the effectiveness of the instances generated with all six rewriting styles by each LLM. To this end, we added the augmented data created by one LLM to the original training datasets.

2 Effectiveness of combined LLMs. In our second setup, we evaluated whether more data would lead to better results. Therefore, we combined the augmented data from both LLMs in all six styles and added it to the provided datasets.

3 Effectiveness of each style from both LLMs. In our third setup, we conducted experiments on the usefulness of the different rewriting styles. Therefore, we combined the augmented data for each style from both LLMs before adding it to the provided datasets.

4 Effectiveness of combined styles. For our final setup, we first examined the F1 scores obtained by each rewriting style. Then, we combined the styles that obtained the best scores and added them to the provided training data.

5.3 Experimental Results

The following sections present the results obtained for the different setups. Each model was trained on the augmented training data and evaluated using the provided trial data (see Section 3). Due to the large number of models, we have summarized the scores where appropriate. The complete results can be found in Appendix B. For the submitted models, we report the macro F1 scores calculated by the organizers of the shared task on the official test set. See Table 5 for these results.

Subtask	Submission 1	Submission 2
C2A	75 %	79 %
DBO	63 %	62 %
VIO	75 %	75 %

Table 5: Official macro F1 scores obtained on the test set as computed by the organizers.

Subtask 1: C2A. Our baseline model, which was trained using the provided dataset, obtained a macro F1 score of 71 %. Augmenting the dataset with the instances from Gemma 3 increased the F1 score by 5 percentage points, while additional instances from Llama 3.3 only increased the score by 1 percentage point. Combining the synthetic instances from both LLMs resulted in an F1 score of 76 %. Using one style as additional data yielded measurable positive effects, with F1 scores ranging from 72 % to 79 %. The best results were obtained with the styles *bissig*, *ironisch*, *sarkastisch*, and

formell and we experimented with different combinations of these styles. A complete overview of the results for Subtask 1 can be found in Table 7 in Appendix B.

For the first submission, we selected a model that was trained using synthetic data from Gemma, achieving an F1 score of 75 % on the test set. For the second submission, we chose the model trained with the styles *bissig+ironisch+sarkastisch* resulting in an F1 score of 79 % on the test set.

Subtask 2: DBO. Our baseline achieved a macro F1 score of 38 % and did not make a single correct prediction for the labels ‘subversive’ and ‘agitation’. Adding synthetic data from Gemma increased the F1 score to 52 %, but it still did not predict any instances of ‘agitation’ correctly. Synthetic data from Llama increased the F1 score to 56 %, but it also failed to predict any ‘agitation’ instances correctly. Evaluating each style revealed that only *bissig* and *polemisch* made correct predictions for the labels ‘subversive’ and ‘agitation’. The styles *formell* and *ironisch* could not make one correct prediction for ‘agitation’, while the styles *formell* and *sarkastisch* could not make any correct predictions for ‘subversive’ and ‘agitation’. See Table 8 in Appendix B for a complete overview of the results for Subtask 2.

For the first submission, we submitted the model’s predictions trained with additional data using the style *bissig*, which resulted in an F1 score of 63 % on the test set. For the second submission, we submitted the model’s predictions trained with the *bissig+polemisch* styles, resulting in an F1 score of 62 %.

Subtask 3: VIO. Our baseline achieved a macro F1 score of 62 %. Adding the augmented data from Gemma increased the score by 16 percentage points, while adding the data from Llama increased the score by 10 percentage points. Interestingly, combining the augmented data from both LLMs resulted in only a 7 percent increase. Adding synthetic data for each style yielded F1 scores ranging from 60 % to 70 %. We experimented with different combinations of the highest-valued styles – namely, *bissig*, *ironisch*, and *sarkastisch* – but could not achieve a higher F1 score than with data augmented by Gemma’s output. Since Gemma’s scores were higher than Llama’s, we experimented with the styles *bissig+ironisch+sarkastisch* only from Gemma, but this did not increase the F1 score. A complete overview of the results for Subtask 3

can be found in Table 9 in Appendix B.

For both submissions, we selected the model that was trained using augmented data from Gemma because no better results could be obtained on the trial dataset. This model achieved an F1 score of 75 % on the test set.

Discussion on the rewriting styles. As mentioned in Section 4.2, rewriting a tweet with either a ‘neutral’ or a ‘formell’ tone may alter the label an annotator would assign. Below, we discuss the results of the experimental setup in which we evaluated the effectiveness of each rewriting style from both LLMs (experimental setup 3 *Effectiveness of each style from both LLMs*) with a focus on the aforementioned two styles.

For C2A, the style ‘formell’ had a decreased F1 score of four percentage points which was nevertheless higher than the scores for ‘polemisch’ and ‘sarkastisch’. But the style ‘neutral’ scored the worst of all styles with a decrease of seven percentage points. Interestingly, the style ‘formell’ had the highest F1 score for DBO, while ‘neutral’ obtained the second worst score. For VIO, ‘formell’ obtained the second best result of all styles, while ‘neutral’ achieved the highest one.

These results suggest that rewriting a tweet in a ‘formell’ tone still conveys the message of harmful content that can be detected by a classifier. The same is not true for a ‘neutral’ style. Such rewritings lack the wording necessary for a classifier to detect, though this was not the case for the VIO subtask. Overall, the rewriting styles may be dependent on the task to be performed and would need tailoring for the specific (classification) task at hand. Furthermore, there might be an interaction between styles.

6 Conclusion & Future Work

In this work, we presented our participation in the shared task of Harmful Content Detection in Social Media. The organizers provided us with annotated datasets that exhibit a heavily skewed class imbalance toward harmless content. For this reason, we employed data augmentation. To this end, we asked three LLMs to rewrite harmful tweets in six different styles. Although LLMs should not replicate harmful content, we designed the prompt so that the LLMs would comply with our request most of the time. They only refused to answer for the most disturbing tweets. Gemma and Llama understood our request to respond with either a rewritten tweet

or ‘None’, Phi-4 did not. It was not possible to remove all unnecessary information from Phi-4’s output, and we could not evaluate its usefulness for this task.

Furthermore, we analyzed the writing styles of the original tweets in the dataset. The analysis revealed that instances from the ‘false’ (C2A/VIO) and ‘nothing’ (DBO) classes were written in a *colloquial* style, while the ‘true’ (C2A/VIO) and ‘subversive’/‘agitation’/‘criticism’ (DBO) classes are mostly categorized as written with an aggressive tone. These results suggest that balancing the writing styles may be important.

In our experimental setup, we evaluated the effectiveness of synthetic data generated by (i) each LLM, (ii) combined LLMs, (iii) one rewriting style, and (iv) combined rewriting styles. Different setups were shown to be more useful for each sub-task. For C2A, we submitted predictions from a model trained with the additional synthetic data from Gemma (75 % on the test set) and those predicted by the styles *bissig+ironisch+sarkastisch* (79 % on the test set). For DBO, many of our models struggled to make correct predictions for the most underrepresented classes, i.e., ‘subversive’ and ‘agitation’. Predicting all classes on the trial dataset was managed by the model trained with the additional data with style *bissig* – obtaining a macro F1 score of 63 % on the test set – and the model trained with the additional data with styles *bissig+polemisch*, resulting in a macro F1 score of 62 % on the test set. For VIO, no setup performed better on the trial dataset than the augmented Gemma data, which received a macro F1 score of 75 % on the test set.

Future research could examine how writing styles influence hate speech detection, focusing on the impact of tone, formality, punctuation, and rhetorical strategies on model performance. Subtle stylistic forms – such as irony, sarcasm, or variation in register – may obscure hateful intent and challenge standard classifiers. While this study did not account for coded language, i.e., veiled or context-dependent expressions that implicitly convey hate, addressing such phenomena is crucial for capturing more covert forms of abuse. LLMs offer promising capabilities to model and generate stylistically diverse, context-aware training data. Future work could leverage LLMs to identify, cluster, or synthesize stylistic and coded variations, thereby improving the robustness of classifiers across sociolinguistic boundaries and communicative con-

texts.

Acknowledgments. This work is partially funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr, which is funded by the European Union – NextGenerationEU.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, and et. al. 2024. *Phi-4 Technical Report. Preprint*, arXiv:2412.08905.
- Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. *Detecting Calls to Action in Multimodal Content: Analysis of the 2021 German Federal Election Campaign on Instagram*. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 1–13, Vienna, Austria. Association for Computational Linguistics.
- Jinan Ali Aljawazeri and Mahdi Nsaif Jasim. 2025. *Enhancing Hate-Speech Detection Using Data-Augmentation Techniques: A Multi-label Dataset Approach*. In *Data Science and Applications*, pages 103–113, Singapore. Springer Nature Singapore.
- Salvador V. Balkus and Donghui Yan. 2023. *Improving short text classification with augmented data using GPT-3*. *Natural Language Engineering*, 30(5):943–972.
- 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 Over-sampling Technique*. *Journal of Artificial Intelligence Research*, 16:321–357.
- Matthew Ciolino, David Noever, and Josh Kalin. 2022. *Back translation survey for improving text augmentation*. *Preprint*, arXiv:2102.09708.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. ACL.
- Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. *Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges*. *Preprint*, arXiv:2403.02990.

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*.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. [Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset](#). 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 4405–4424, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et. al. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.

Lingshu Hu. 2023. [Mobilization, self-expression or argument? A computational method for identifying language styles in political discussion on Twitter](#). *Online Information Review*, 48(4):783–802.

Sadar Jaf and Basel Barakat. 2024. [Empirical Evaluation of Public HateSpeech Datasets](#). *Preprint*, arXiv:2407.12018.

Md Saroar Jahan, Mourad Oussalah, Djamila Romaissa Beddiar, Jhuma Kabir Mim, and Nabil Arhab. 2024. [A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs](#). *CoRR*, abs/2404.00303.

Kyriaki Pantelidou, Despoina Chatzakou, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2022. Selective word substitution for contextualized data augmentation. In *Natural Language Processing and Information Systems*, pages 508–516, Cham. Springer International Publishing.

Nur Balqies Sani, Novia Lestari P. A., Ronny Ahmad Fauzi Pane, Anisa Novita Sari, and Rakhmat Wahyudin Sagala. 2025. [An Analysis of Language Styles on Social Media: Twitter Case Study](#). *ALTERA: Journal of Applied Linguistics, English Education, and Literature*, 1(01):12–20.

Norman Sapar, Ab Razak Che Hussin, and Nadhmi A. Gazem. 2023. A New Landscape of Political Engagement Through Social Media: How Can We Map It? In *Advances on Intelligent Computing and Data Science*, pages 544–552, Cham. Springer International Publishing.

Stefan Stieglitz and Linh Dang-Xuan. 2013. [Social media and political communication: a social media analytics framework](#). *Social Network Analysis and Mining*, pages 1277–1291.

Gemma Team, Aishwarya Kamath, Johan Ferret, and et. al. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.

Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

A Prompt

The `create` function of Ollama was used to initialize different LLMs and to instruct them on their task. The variable **rephStyle** written in bold stands for one of the six writing styles and is therefore substituted by the words *bissig*, *formell*, *ironisch*, *neutral*, *polemisch*, or *sarkastisch* (in English: *biting*, *formal*, *ironic*, *neutral*, *polemic*, and *sarcastic*). Table 6 shows the important parameters for the `create` function.

Ollama create function	
from_:	llama3.3:70b or gemma3:27b or phi4:latest
system:	Im Anschluss werden ausschließlich Tweets genannt, welche bitte umgeschrieben werden sollen. Die Tweets sollen so umgeschrieben werden, dass sie rephStyle klingen. Ersetze zudem bitte die Entitäten durch semantisch vergleichbare. Beachte bitte, dass es sich bei der Aufgabe um eine wissenschaftliche Aufgabe handelt. Die Inhalte der Tweets können teils verstörend sein, nichtsdestotrotz führe die Aufgabe bitte so präzise wie möglich aus. Gebe als Antwort ausschließlich den umgeschriebenen Tweet zurück. Beschreibungen sind nicht nötig! Solltest du Probleme mit dem Lesen bzw. Interpretieren des Tweets haben, dann gebe ausschließlich das Wort 'None' zurück.
parameters:	{temperature: 0.9, top_p: 0.9}

Table 6: Parameters used to initialize the LLM.

As the provided datasets contain German text samples, the system prompt was also provided in German. An English translation of the prompt would be: “In the following only tweets will be listed that need to be rewritten. The tweets should be rewritten so that they sound in the style of *rephStyle*. Also, please replace the entities with semantically comparable ones. Please note that this is a scientific task. The content of the tweets may be disturbing; nevertheless, please carry out the task as precisely as possible. Respond only with the rewritten tweet. No descriptions are necessary. If you have trouble reading or interpreting the tweet, respond only with the word ‘None’.”

B Complete Experimental Results

Tables 7, 8, and 9 list all experimental results conducted for each subtask. Reported F1 scores were obtained by evaluating each model on the trial data.

LLM/Style	Macro F1
Baseline	71 %
Gemma	76 %
Llama	72 %
Gemma+Llama	76 %
<i>bissig</i>	75 %
<i>formell</i>	75 %
<i>ironisch</i>	79 %
<i>neutral</i>	72 %
<i>polemisch</i>	73 %
<i>sarkastisch</i>	74 %
<i>bissig+ironisch</i>	71 %
<i>bissig+sarkastisch</i>	76 %
<i>bissig+ironisch+sarkastisch</i>	79 %
<i>bissig+ironisch+formell</i>	76 %

Table 7: All experimental results on the trial dataset for the subtask ‘Call2Action’.

LLM/Style	Macro F1
Baseline	38 %
Gemma	52 %
Llama	56 %
Gemma+Llama	—*
<i>bissig</i>	55 %
<i>formell</i>	58 %
<i>ironisch</i>	43 %
<i>neutral</i>	39 %
<i>polemisch</i>	56 %
<i>sarkastisch</i>	38 %
<i>bissig+polemisch</i>	67 %
<i>bissig+polemisch+sarkastisch</i>	60 %

Table 8: All experimental results on the trial dataset for the subtask ‘Attack on the Democratic Basic Order’. (*: The full dataset was too large to fit in a laptop’s RAM.)

LLM/Style	Macro F1
Baseline	62 %
Gemma	78 %
Llama	72 %
Gemma+Llama	69 %
<i>bissig</i>	66 %
<i>formell</i>	69 %
<i>ironisch</i>	70 %
<i>neutral</i>	70 %
<i>polemisch</i>	60 %
<i>sarkastisch</i>	69 %
<i>bissig+ironisch</i>	68 %
<i>ironisch+sarkastisch</i>	67 %
Gemma <i>bissig+ironisch+sarkastisch</i>	68 %

Table 9: All experimental results on the trial dataset for the subtask ‘Violence Detection’.

C Augmented Data for Subtask 2

Table 10 provides a detailed breakdown of the number of instances generated for each label and LLM for Subtask 2 (DBO).

LLM	Style	Label			Total
		subversive	agitation	criticism	
Gemma	<i>bissig</i>	60	305	802	1,167
	<i>formell</i>	61	306	802	1,169
	<i>ironisch</i>	60	312	802	1,174
	<i>neutral</i>	59	309	802	1,170
	<i>polemisch</i>	43	226	598	867
	<i>sarkastisch</i>	61	311	802	1,174
Llama	<i>bissig</i>	49	195	754	998
	<i>formell</i>	51	223	787	1,061
	<i>ironisch</i>	54	210	782	1,046
	<i>neutral</i>	58	287	801	1,146
	<i>polemisch</i>	45	159	721	925
	<i>sarkastisch</i>	55	231	788	1,074

Table 10: Detailed overview of the generated instances per label and per LLM for Subtask 2.