Overview of the GermEval 2025 Shared Task on Harmful Content Detection

Jenny Felser and Michael Spranger
Mittweida University of Applied Sciences
Technikumplatz 17
09648 Mittweida

Melanie Siegel
Darmstadt University of Applied Sciences
Schöfferstr. 3
64295 Darmstadt

Abstract

Social media is not only a channel for the appropriate sharing of personal opinions and enriching discussions, but also facilitates the dissemination of inappropriate and aggressive statements. Those are especially concerning when they actively incite harmful actions such as violence or attacks on the government. Against this background, this paper presents the GermEval Shared Task for Harmful Content Detection, which addresses three subtasks that have been largely neglected in previous competitions and research projects: the detection of 1) calls to action, 2) attacks against the liberal democratic basic order and 3) violence-supporting statements. For this pilot task, 11,551 tweets from a German Twitter network belonging to an extremist group were annotated. A total of eleven teams participated in at least one of the three subtasks, with nine teams submitting a system paper. Overall, macro-average F_1-scores of up to 87% were achieved for subtasks 1 and 3 and up to 71% for subtask 2.

Content warning: We show illustrative examples of harmful content.

1 Introduction

Social media allows users to express their opinions, views and perceptions freely. However, the downside of freedom of expression is that increasingly toxic, hateful and aggressive content is being spread. That does not only include the mere expression of hurtful comments but goes so far as to incite other people to hate or even call for violent action. One of the best-known examples is the murder of Walter Lübcke, in which hate comments were posted by those involved in the crime in advance, which also contained calls for murder and incitement to overthrow the government violently (Thorwarth, 2023). From a prevention perspective, it is therefore essential to recognise incitement, advocacy of violence, and even incitement of violence in good time to take countermeasures. In this way, for example, moderators of online social networks may be assisted in removing harmful content.

Another focus in connection with current political developments is (verbal) attacks on the liberal Democratic Basic Order (DBO) by subversive movements. In 2024, the German Federal Office for the Protection of the Constitution warned that right-wing extremists, for example, are using microblogging services such as X (formerly Twitter) to spread propaganda, mobilise supporters and even announce attacks (Bundesamt für Verfassungsschutz, 2024). In this context, the detection of harmful content such as incitement to violence and subversive tendencies is not only essential from a preventive perspective, but also supports law enforcement authorities in analysing mass communication in social networks during criminal investigations.

Against this background, we are introducing the GermEval Shared Task on Harmful Content Detection. This shared task aims to initiate and promote research into the detection of three particularly worrying types of harmful content in German social media posts: 1) so-called calls to action, 2) attacks on the DBO and 3) worryingly positive statements about violence. For this purpose, three new German-language Twitter datasets were created, each based on the annotation of more than 11,500 tweets from the network of a right-wing extremist group. The eleven teams participating in the competition made an initial contribution to addressing the three tasks, which have hardly been tackled to date.

The remainder of this overview paper is structured as follows: section 2 provides an overview of related work, and section 3 presents the three subtasks addressed. The dataset and its creation are described in section 4. Section 5 presents the evaluation and the baseline systems, before section 6 discusses the approaches and results of the participants. Finally, we conclude in section 7.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 306–319, 2025

Objective/ task	Competition	Language
offensive language detection	GermEval 2018 (Wiegand et al., 2018)	German
	GermEval 2019 (Struß et al., 2019)	German
	GermEval 2021 (Risch et al., 2021)	German
hate speech detection	HASOC 2020 (Mandl et al., 2020)	German, English, Hindi
	HASOC 2022 (Satapara et al., 2022)	German, Hinglish
sexism and/or misogyny detection	GermEval 2024 (Gross et al., 2024)	German
	EXIST 2025	English, Spanish
classification of conspiracy theories	PAN 2024 (Korenčić et al., 2024)	English, Spanish
political bias detection	SemEval 2019 (Kiesel et al., 2019)	English
	CheckThat! 2023 (Barrón-Cedeño et al., 2023)	English
classification of violence incitements	BLP-2023 (Saha et al., 2023)	Bangla

Table 1: Overview of previous shared tasks in the area of detecting harmful, hartful or problematic content

2 Related Work

Harmful content in general has already been addressed by numerous shared tasks, with some examples listed in Table 1. However, earlier GermEval shared tasks focused mainly on the detection of offensive (toxic) language (e.g. Wiegand et al., 2018). Closely related to this task is the detection of hate speech and sexism, for which German-language resources have also been developed as part of shared tasks (e.g. Mandl et al., 2020). However, the focus has not yet been on detecting calls to action, attacks on the DBO, and incitement to violence:

Detection of calls to action. To our knowledge, calls to action detection has not been addressed by any previous shared tasks, and only a few datasets have been created in a small number of research projects in this area (e.g. Siskou et al., 2022). However, those did not focus on the German language and only considered a specific type of call to action, such as calls for harmful actions against a particular ethnic group (Pérez et al., 2023) or for the mobilisation of voters during an election in Spanish social media (Siskou et al., 2022). One of the few studies based on a German-language dataset of 1,388 Instagram posts aimed to identify calls to action in the context of the 2021 German federal elections (Achmann-Denkler et al., 2024). In contrast, subtask 1 of our shared task encompasses not only the mobilisation of voters, but also the identification of all calls to action, including those for rallies, demonstrations and protests.

The identification of calls for protests and gatherings has previously been investigated by Rogers et al. (2019), for example, using Russian tweets. However, as the authors themselves point out, the problem is that authoritarian regimes could abuse such methods to track down and suppress legitimate protests. The motivation behind the first subtask of the shared task, by contrast, is to identify and manually check calls with the express aim of distinguishing legitimate calls for protest from those inciting violent acts, and to support security force decision-makers in planning resources for potentially dangerous situations.

Detection of attacks on the DBO. Furthermore, there have been no shared tasks or other research projects to date that deal with the detection of subversive intentions and further attacks on the DBO. Instead, some shared tasks addressed the classification of conspiracy theories (Korenčić et al., 2024) and political bias (e.g. Kiesel et al., 2019) (see Table 1), which may also be expressed in anti-government tweets attacking the DBO. However, the data used for these shared tasks did not include German texts and, in some cases, consisted of newspaper articles (Kiesel et al., 2019; Barrón-Cedeño et al., 2023), which differ greatly from social media posts in terms of language and length.

In addition, research projects have focused on identifying verbal accusations against the government (e.g. Lemmens et al., 2022; Corral et al.,

2024), such as allegations of hypocrisy against politicians (Corral et al., 2024). Furthermore, efforts have been made to identify extremist views (e.g. Gaikwad et al., 2022). However, this work differs from our second subtask in that the latter involves deciding whether concrete actions against the government are planned, which is particularly important in terms of timely preventive measures.

Violence detection The only shared task dedicated to the announcement of violence uses YouTube comments in Bangla as its data source, with transformer models explicitly developed for Bangla achieving the best results (Saha et al., 2023). Another interesting study aimed to use a convolutional neural network to detect calls for violence in Urdu tweets (Khan et al., 2024). In addition to the fact that these studies did not address the German language, the third subtask stands out in that it is not only a matter of recognising incitements of violence, but also other types of questionable statements about violence, such as the glorification of violence. The task of recognising whether violence is being discussed in social media has already been addressed by Cano Basave et al. (2013) for English. However, they did not evaluate the authors' attitudes towards violent events.

In summary, the new shared task makes an essential contribution to research into the detection of harmful content, particularly concerning the German language, by providing new data sets.

3 Task Description

Participants were allowed to take part in one, two or all three subtasks, with up to three runs per subtask. All subtasks were designed as open tracks, where participants can use external datasets that have been annotated for the subtasks or related tasks (e.g. Gross et al., 2024). This decision was made to encourage data enrichment strategies that may be promising given the imbalance of the dataset described in section 4. Specifically, the shared task comprised the following subtasks:

Subtask 1: Detection of Calls to Action The first subtask was a binary decision as to whether a tweet contained a call to action (TRUE) or not (FALSE). According to the Oxford Dictionary definition, a call to action is defined as a command or request to perform a specific action or behaviour (Oxford University Press, 2024). This action may be criminally relevant if it is incited or encouraged, but this is not necessarily the case. It could also be a call for a legitimate demonstration. Examples[1] of calls to action are:

- AWO abschaffen. (engl. Abolish AWO)
- Ihr müsst schnell handeln ,sonst ist es zu spät.Ansonsten haben die uns schon überrannt. (engl. You have to act quickly, otherwise it will be too late. Otherwise, they will have overrun us.)

Subtask 2: Classification of attacks on the DBO The second subtask consists of a fine-grained classification in four categories (subversive, agitation, criticism, nothing) concerning various harmless and harmful forms of *attacks on the DBO of the Federal Republic of Germany*:

- **subversive:** Based on the definition in the Duden dictionary, a subversive tweet expresses the desire to overthrow the current government or state order violently, e.g. through militant action or paralysing the power grid (Duden, 2025).
- **agitation:** The tweet conveys inflammatory intentions against the DBO, such as the dissemination of propaganda by unconstitutional and terrorist organisations as defined in Section 86 of the German Criminal Code (StGB) or the defacement of state symbols such as the flag of the Federal Republic of Germany (§ 90a StGB).
- **criticism :** The tweet contains legal but negative criticism of the government, civil servants, government employees, authorities or political parties.
- **nothing:** The tweet does not contain any attacks on the DBO. Neutral or positive statements about government decisions also fall into this category.

Examples of tweets from each category in the training data set can be found in Table 2.

Subtask 3: Violence Detection The third task is determining whether the tweets express a disturbingly positive attitude towards violence (TRUE,

[1]All examples shown in this paper are taken from the newly created training dataset. Spelling mistakes were not corrected. An English translation is provided for each tweet.

Label	Example
subversive	...dieser Wahnsinn muss aufhören...weg mit dieser VolksVerräterVerrücktenRegierung...!!!
	...this madness must stop...away with this traitorous, crazy government...!!!
agitation	was für gesetze machen die eigentlich ,EU ??? ,kassieren 20.000 im Monat und lassen uns untergehen , EUROPA - Asoziale Politik .
	What kind of laws are they making, EU??? They earn 20,000 a month and let us go under, EUROPE – anti-social politics.
criticism	Die Polizei wird nur unnötig unter Druck gesetzt, wenn ständig irgendwelche Gegendemos genehmigt werden.
	The police are only put under unnecessary pressure if counter-demonstrations are constantly being authorised.
nothing	Sachsen ist Deutschland und Deutschland ist Sachsen
	Saxony is Germany, and Germany is Saxony.

Table 2: Examples of tweets from the categories *subversive*, *agitation*, *criticism* and *nothing* of the second subtask, each with an English translation.

FALSE). Such disturbing statements include trivialisation, support, glorification, incitement or announcement of violent acts, such as:

- *jo TOT gefallen sie mir auch am besten :)* (engl. *Yes, I like them best dead too :)*)
- *Der Schießbefehl muss unverzüglich gegeben werden* (engl. *The order to shoot must be given immediately.*)

4 Data Set

The primary source of data was previously collected, unannotated German tweets, which were annotated for the three subtasks. Underrepresented categories were supplemented with tweets from existing German datasets on hate speech detection (Demus et al., 2022; Kums et al., 2025).

4.1 Initial data collection

The bulk of the training and test data for the three subtasks is based on public German Twitter posts and comments in the context of a group considered as right-wing extremist. This data source was chosen because the Shared Task focuses on the German language, and content from the group's environment was expected to frequently serve mobilisation purposes, including calls for high-risk actions and actions against the government. The posts and comments were initially collected by members of Mittweida University of Applied Sciences from December 2014 to July 2016. For the annotation, a sample of posts and corresponding comments was selected from over 800,000 tweets, linking them via their respective post IDs. The sample comprised a total of 11,551 tweets to be annotated, including 54 posts.

To ensure anonymity, user mentions were semi-automatically replaced by [@PRE] for mentions of the press, [@POL] for the police, [@GRP] for groups and [@IND] for individuals. In this way, the information about who the tweet is addressed to can still be used to develop the models.

4.2 Data annotation

Members of Mittweida University of Applied Sciences annotated the 11,551 tweets. A total of 88 people were involved in the annotation process. The dataset was split into samples of around 500 tweets, each of which was annotated independently by four annotators, i.e., each annotator labelled approximately 500 tweets. The annotators were either students or research assistants in digital forensics and had expertise in identifying the characteristics of harmful textual content. They had confirmed their consent to the publication of the data in writing. The annotators were provided with comprehensive annotation guidelines and allowed to use external resources, such as dictionaries for unknown terms, in the comments.

Subtask 3 was originally annotated at a fine granularity, categorising worrying positive statements about violence into five subtypes (propensity, call to violence, support, glorification, other forms). However, since initial analyses showed that some categories were severely underrepresented, this task was converted into a binary classification.

For the binary detection of calls to action, the annotators could also indicate that they were unable to judge this. In addition to the annotations required for this shared task, the data set was annotated for hate speech, toxicity, target and emotions to support the classification tasks.

The final data sets for each subtask contained all tweets on which a majority decision was reached among the four annotators. In the case of the first subtask, tweets were also sorted out where the majority of annotators stated that they were unable to evaluate the tweet. The resulting datasets comprised 9,822 tweets for the first subtask, 9,307 for the second subtask, and 10,933 for the third subtask. The Fleiss kappa coefficient (Fleiss, 1971) was calculated for these data sets. The agreement between the four annotators was $\kappa = 0.391$ for the first subtask, $\kappa = 0.416$ for the second subtask, and $\kappa = 0.376$ for the third subtask. According to Landis and Koch (1977), these results correspond to a fair agreement for subtasks 1 and 3 and a moderate agreement for subtask 2, with the values for subtasks 1 and 2 close to the threshold of above 0.4 for moderate agreement. The relatively low level of agreement, even when compared to annotations of other types of harmful content, such as offensive language (e.g. Wiegand et al., 2018; Demus et al., 2022), can be attributed not only to the lack of annotation experience of the majority of annotators, but also to the complexity of the tasks and subjective differences in assessment. For example, when deciding at what point statements relating to violence should be classified as problematic. The following tweet, for instance, was assigned to the violence category by the majority of annotators.

> Baut das Ding. Am besten aus Asbest...
> (engl. Build that thing. Preferably out of asbestos...)

Although it seems likely that the tweet expresses a desire to cause someone physical harm, without knowing what kind of building is being referred to, it is difficult to assess whether there is a connection to violence.

4.3 Data augmentation

An initial analysis of the data sets revealed that the relevant categories were significantly underrepresented, particularly in subtasks 2 and 3. For subtask 2, for example, the data set contained only 29 subversive posts and 49 tweets with inflammatory content. In subtask 3, the proportion of violence-related statements was only 5.67%. Therefore, the data sets for the two subtasks were enriched as follows:

Subtask 2: For subtask 2, another annotator identified comments from the two German-language X (formely Twitter) datasets Detox (Demus et al., 2022) and the dataset described by Kums et al. (2025) from the field of hate speech that could be assigned to the four subclasses. The focus was particularly on enriching the two categories, subversive and agitation. Overall, the original dataset for the second subtask was expanded by 1,341 tweets.

Subtask 3: For the third subtask, the TRUE category (violence-related statements) was enriched using the hate speech dataset by Kums et al. (2025). In this dataset, the comments were manually classified relating to different offences under the Criminal Code. Only the 211 comments assigned to the category *Public Incitement to Commit Crimes and Disturbing the Public Peace*[2] were considered for enrichment, which were most likely to contain problematic, violence-related statements. These comments were re-annotated by a shared task organiser. 185 comments could be added to the violence category. The remaining comments were highly racist but did not directly refer to violence and were therefore not relevant to subtask 3.

Details about the composition of the data sets after augmentation are listed in subsection A.1.

4.4 Data splitting

The resulting data sets for the three subtasks comprised approximately 9,000 to 11,000 tweets, with a high degree of overlap of 7,708 tweets. The data sets for each subtask were split into training and test data using stratified sampling in a ratio of 70:30. In addition, a stratified sample of approximately 1,000 tweets was taken from the annotated training data and used as trial data. Those were published prior to the training data so that participants could familiarise themselves with the data format. All data sets were made available as semicolon-separated

[2]This category includes offences that are punishable under §111 StGB, §126 StGB, §130 StGB, §131 StGB, §140 StGB (Kums et al., 2025).

CSV files on GitHub[3]. For each tweet, only the ID was provided in addition to its text content; no other metadata was included. An example of the data format can be found in subsection A.2.

4.5 Description of the resulting data sets

Class distribution The size and class distribution of the training and test data for the three subtasks are given in Table 3. All data sets were highly unbalanced, with a low proportion of relevant and harmful content. This observation is consistent with the class distribution of data sets that have been annotated for the detection of other hurtful content, such as offensive language (Struß et al., 2019). The problem of imbalance is most evident in the fine-grained classification in subtask 2 for detecting attacks on the DBO (see second row in Table 3). The category subversive is particularly underrepresented, which presumably – and in a way fortunately – also reflects the actual, rare occurrence of such tweets in social media.

Linguistic properties Aside from the considerable imbalance in the data sets, another challenge is the linguistic characteristics of the tweets. To illustrate this, Table 4 presents the average usage of some token types in the tweets of the training data sets for the three subtasks.

Property	Subtask (1)	(2)	(3)
⌀ tokens	29.35	27.44	29.64
⌀ chars	158.40	150.20	161.50
⌀ emoticons	0.19	0.19	0.20
⌀ capslock	0.33	0.33	0.36
⌀ hashtags	0.02	0.09	0.03
⌀ mentions	0.01	0.13	0.01

Table 4: Statistical description of the trainining data sets for (1) call to action detection, (2) detection of attacks on the DBO, (3) detection of worrying violence-related statements.

The tweets all have a small number of tokens (words, but also other units such as punctuation marks, symbols, etc.), which poses a challenge for text classification (Alsmadi and Gan, 2019). In addition, they contain components typical of social media, such as emoticons and caps lock[4]. The differences between the datasets were mostly minor, which can be explained by the high degree of overlap. However, the dataset from subtask 2 contains more hashtags and mentions than the other two datasets. A detailed analysis of the hashtags and mentions in the individual categories would go beyond the scope of this paper. An initial analysis showed that @IND mentions were most common in the dataset for subtask 2, mostly in response to another tweet and using hashtags referring to politicians, for instance *#Merkel* and *#MerkelMussWeg* or using the Hitler salute *#SiegHeil*.

5 Evaluation metrics and baselines

As with previous shared tasks by GermEval (e.g. Gross et al., 2024; Schomacker et al., 2024), the Codabench platform[5] was used for evaluation. The competition submissions were ranked using the macro-average F_1-score. This evaluation metric is an obvious choice, as all classes are treated equally instead of giving greater importance to the detection of harmless, overrepresented content. Accordingly, the macro-average F_1-score has already been used as the primary evaluation metric in previous shared tasks for detecting harmful content (Wiegand et al., 2018; Struß et al., 2019; Risch et al., 2021). Apart from that, the macro-average precision and recall were calculated.

A simple baseline system was developed for each subtask (classical methods, embeddings, large language models (LLMs)), and the code was published on GitHub[6]. The baselines represent obvious approaches and are intended to encourage participants to try out alternative, more innovative methods:

Subtask 1: For the first subtask, a gradient boosting classifier (Friedman, 2001) was used, whereby the tweets were represented as dense vectors using a pre-trained Sentence BERT model (Reimers and Gurevych, 2019). As an additional feature, the polarity score of the tweet was used, which was determined with a lexicon-based method using the TextBlob library (Loria, 2018). The underrepresentation of calls to action was addressed by random

[3] https://github.com/Communication-Forensics-Lab/harmful-content-detection/tree/main/data

[4] words written entirely in capital letters

[5] https://www.codabench.org/competitions/4963/

[6] https://github.com/Communication-Forensics-Lab/harmful-content-detection/tree/main/baseline

Subtask	Class Label	Training Data Freq	Training Data %	Test Data Freq	Test Data %
(1) Calls to action	TRUE	663	9.69	289	9.69
	FALSE	6177	90.31	2693	90.31
	Total	6840	100.00	2982	100.00
(2) Attacks on the DBO	subversive	60	0.80	25	0.78
	agitation	313	4.20	134	4.20
	criticism	804	10.79	345	10.80
	nothing	6277	84.21	2690	84.22
	Total	7454	100.00	3194	100.00
(3) Violence-related statements	TRUE	564	7.25	241	7.23
	FALSE	7219	92.75	3094	92.77
	Total	7783	100.00	3335	100.00

Table 3: Class distribution of training and test data in the three subtasks.

undersampling, after which the two categories contained the same number of tweets.

Subtask 2: A simple classification approach based on a linear Support Vector Machine (SVM) (Chang and Lin, 2011) with TF-IDF-weighted bag-of-phrases (unigrams and bigrams) was chosen as the baseline system. The vocabulary was limited to the 5000 most frequent phrases in the training data set. The class imbalance was taken into account by a cost-sensitive (weighted) SVM, in which misclassifications of instances from underrepresented classes are more penalised.

Subtask 3: For the third subtask, classification was performed by the generative open source large language model Qwen2.5 (Bai et al., 2023) with 32 billion parameters. In-context learning was chosen as the prompting strategy, which has already proven promising in the detection of related harmful content such as hate speech (Assis et al., 2024; Sahin et al., 2023). Details on prompt engineering can be found in Appendix A.

6 Results

A total of eleven teams participated in at least one subtask of the shared task, with six teams participating in all three subtasks. A summarised statistic of the macro-average F_1-score achieved in the individual tasks can be found in Table 5. As the table shows, the results achieved by the individual teams vary greatly. In the second task in particular, the range of scores was approximately 41%. The macro-average F_1-score for this subtask also showed the highest standard deviation, indicating that the effectiveness of the systems varied greatly. On average, higher values were achieved for subtasks 1 and 3 than for subtask 2. Possible causes can be found in the strong imbalance of the training data set for subtask 2.

Subtask 1: Participation was highest in the detection of calls to action: a total of 20 valid runs were submitted by nine teams, with the results listed in Table 6. In this subtask, F_1 scores of up to 86.98% were achieved, with the two highest results being obtained by *SuperGLEBer*. This team fine-tuned the ModernGBERT 1B model (Ehrmanntraut et al., 2025) for their third run, the best-performing run, and an LLäMmlein 7B model (Pfister et al., 2025) for their second run. In addition, all but one run outperformed the baseline, the gradient boosting classifier.

Subtask 2: We received a total of 16 valid runs from seven teams for the classification of attacks on the DBO, with the results shown in Table 7. The run submitted by the *munich_z* team during the competition phase was invalid. After the end of the competition, the team submitted three valid runs, which are marked in grey in Table 7 for assessment of system performance, but were not included in the summary statistics in Table 5. Considering the results, it is apparent that the macro-average F_1 values achieved were lower than those of the other

Subtask	# Teams	# Valid Runs	Min	Max	Median	Mean	SD
(1) call to action	9	20	53.93	86.98	78.81	76.27	8.94
(2) DBO	7	16	29.86	71.22	64.63	59.09	14.56
(3) violence	8	17	67.12	86.76	79.10	77.19	5.64

Table 5: Summary statistics of the macro-average F_1-scores achieved in the three subtasks.

two subtasks, with the team *nymera* achieving the best value of 71.22% with an ensemble of three different finely tuned BERT variants. Six teams were able to exceed the SVM baseline.

Subtask 3: The results achieved for the detection of worrying, violence-related statements, are illustrated in Table 8. A total of 17 valid runs were submitted by eight teams, with the best macro-average F_1-score of 86.76% again achieved by *SuperGLEBer*, this time with the LLäMmlein 7B model (Pfister et al., 2025) (run two by *SuperGLEBer*) outperforming the ModernGBERT 1B (Ehrmanntraut et al., 2025) (run 3) model. Furthermore, almost all teams except one exceeded the baseline using the Qwen2.5 model (Bai et al., 2023).

team	run	P	R	F_1
SuperGLEBer	**3**	**86.70**	**87.26**	**86.98**
SuperGLEBer	2	85.70	85.15	85.42
nymera	2	84.99	82.26	83.56
HSH;-)	2	82.97	84.09	83.52
nymera	3	84.65	81.40	82.92
nymera	1	84.30	80.86	82.46
abullardUR	3	79.36	86.12	82.25
abullardUR	2	81.60	82.57	82.08
HSH;-)	1	80.82	83.08	81.89
FI-CODE	2	81.33	77.70	79.37
LabelLords	1	79.35	77.25	78.25
NLPeace	2	81.44	75.77	78.23
mzb	2	73.77	82.08	77.01
FI-CODE	1	78.73	72.53	75.13
mzb	1	77.19	68.35	71.63
tweetbusters	3	68.02	64.54	66.01
tweetbusters	1	68.57	63.03	65.10
tweetbusters	2	68.57	63.03	65.10
LabelLords	2	62.39	71.60	64.64
baseline	-	*59.88*	*74.75*	*59.13*
NLPeace	1	54.98	61.37	53.93

Table 6: Results of subtask 1: the detection of calls to action. The macro-average of the precision (P), recall (R) and F_1 measures are given respectively.

team	run	P	R	F_1
nymera	**3**	**72.81**	**70.54**	**71.22**
SuperGLEBer	3	74.50	65.78	69.21
nymera	2	72.43	66.81	69.21
TheDBOs	2	72.94	66.02	68.59
TheDBOs	3	66.14	74.93	68.07
TheDBOs	1	68.37	66.48	67.34
HSH;-)	2	68.92	66.46	66.49
nymera	1	67.89	65.49	66.25
SuperGLEBer	2	68.07	59.98	63.01
FI-CODE	1	68.21	63.15	62.77
abullardUR	2	58.30	69.14	62.60
FI-CODE	2	63.82	62.04	62.36
abullardUR	3	53.81	61.08	56.38
baseline	-	*54.75*	*51.35*	*47.44*
munich_z	2	37.42	47.19	38.11
munich_z	1	36.88	55.50	37.41
munich_z	3	34.70	46.01	32.68
mzb	2	35.62	48.58	31.52
mzb	1	33.66	46.08	30.64
mzb	3	32.84	46.51	29.86

Table 7: Results of subtask 2: classification of different attacks on the DBO. The macro-average of the precision (P), recall (R) and F_1 values are specified.

General Conclusions: The following discussion focuses on the approaches of nine of the eleven participating teams for which a system description was available[7]. Eight of these nine teams used a neural (large) language model for at least one of their runs, with fine-tuning of German or multilingual variants of BERT such as $\text{GBERT}_{\text{Base}}$ (Chan et al., 2020) and XLM-RoBERTa (Conneau et al., 2020) on the training data of the shared task being a particularly popular approach (e.g., *FI-Code*, *HSH;-)*). ModernGBERT 1B ranked among the top three performing systems in all three subtasks (see *SuperGLEBer* run 3 in all subtasks). The smaller ModernGBERT 134M (e.g., *abullardUR*) yielded

[7]The systems used by the mzb and LabelLords teams are not covered here.

team	run	P	R	F_1
SuperGLEBer	**2**	**89.02**	**84.80**	**86.76**
SuperGLEBer	3	86.13	82.93	84.44
mzb	2	81.85	81.49	81.67
abullardUR	2	82.16	80.93	81.53
abullardUR	3	81.36	80.64	81.00
nymera	3	87.01	75.96	80.33
HSH;-)	2	84.04	76.74	79.86
nymera	1	82.64	76.42	79.13
nymera	2	88.40	73.82	79.10
HSH;-)	1	80.29	73.43	76.33
NLPeace	2	82.02	72.59	76.32
FI-CODE	1	73.62	75.51	74.52
FI-CODE	2	73.62	75.51	74.52
NLPeace	1	67.38	81.16	71.44
mzb	1	71.79	69.94	70.81
baseline	*1*	*65.32*	*88.36*	*68.97*
LabelLords	2	64.43	73.64	67.38
LabelLords	1	64.36	72.61	67.12

Table 8: Results of subtask 3: Detection of worrying, violent statements. The macro-average of the precision (P), recall (R) and F_1 values are listed.

more modest results, especially in the first two subtasks, illustrating that model size has a crucial influence on performance, alongside training data and architecture. Another promising approach was the use of a soft voting ensemble consisting of various German (Chan et al., 2020), multilingual (Conneau et al., 2020) and English (He et al., 2021) BERT-based models (Conneau et al., 2020) (e.g., *nymera*), even achieving the best result for subtask 2.

Some teams used modern open-source LLMs, including Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Qwen-3 32B (Yang et al., 2025) and LLäMmlein 7B (Pfister et al., 2025) (e.g., *NLPeace*, *munich_z*, *SuperGLEBer*). Remarkably, LLäMmlein achieved second place in subtask 1 and first place in subtask 3, while other LLM-based approaches mostly failed to exceed the baseline (e.g., *NLPeace* in subtask 1, *munich_z* in subtask 2). One explanation is probably that LLäMmlein was trained exclusively on German-language texts. In systematic comparisons of over 30 models, the *SuperGLEBer* team also demonstrated that models optimised for German outperform multilingual LLMs such as Qwen. Furthermore, SuperGLEBer was the only team to fine-tune an LLM, while the other teams (i.e., *NLPeace*, *Munich_z*) used zero-/few-shot prompting. Although they found that choosing suitable prompts improved the results, LLMs not specifically optimised for German still lagged behind the fine-tuned, smaller BERT-based models. Similarly, the LLM-based baseline system for subtask 3 performed poorly.

None of the top ten systems used classic machine learning methods, which were only investigated in isolated cases (e.g. by *NLPeace*, *tweetbusters*). Nevertheless, *tweetbusters* achieved a macro-average F_1-score of 66.01% in subtask 1 using extensive feature engineering and a soft voting ensemble, outperforming a few-shot mixtral system (i.e. *NLPeace*) by approximately 12%, indicating that classical approaches can still be competitive.

Finally, six teams explicitly addressed the underrepresentation of harmful content in the training data through approaches such as re-sampling (e.g., *NLPeace*, *nymera*), suitable loss functions *nymera*, or data augmentation with LLM-generated examples (e.g., *TheDBOs*, *nymera*, *abullardUR*), including style changes, e.g., *FI-Code*. Since these strategies were combined with different classification models, it is not easy to assess which method is particularly effective. Nevertheless, it is noteworthy that *TheDBOs* achieved one of the best results in subtask 2. Their approach involved iteratively adding high-quality synthetic tweets generated with an LLM (Dubey et al., 2024), concluding that the quality of the synthetic examples is particularly crucial for success.

Error analysis: We further investigated which examples in the test dataset were easy or difficult for participants. In the first subtask, 51.5% of all examples were correctly classified in all runs, in the second subtask, only 25.6%, and in the third subtask, 68.9%. This observation highlights apparent differences between the subtasks in terms of 'easily classifiable' tweets.

Furthermore, for all subtasks, there were a few tweets in the test dataset that were misclassified in every submitted run (8 tweets in the first, 35 in the second and 6 in the third subtask). These often prove difficult for humans to classify as well. For example, the following tweet was classified by the majority of annotators as subversive in the second subtask, but it is questionable whether it expresses a desire for serious violent overthrow.

> Was für Fachkräfte, wenn die im eigenem Land arbeitslos sind, einfach lächerlich die Politik hier in Deutschland absetzen und von ihren Sesseln schubsen da oben.

> pfui ? (engl. What kind of skilled workers are they if they are unemployed in their own country, simply dismissing the politics here in Germany as ridiculous and pushing them out of their seats up there? Disgraceful!)

A detailed error analysis of the individual systems is beyond the scope of this overview paper. However, looking at the best runs for each subtask reveals recurring difficulties. Those include implicit or subtly formulated content, which is consistent with the findings of previous shared tasks for detecting harmful content (e.g. Wiegand et al., 2018). For example, the following tweet was annotated as a call to action, but it does not contain any classic syntactic markers of a call to action, such as the imperative form:

> Im Interesse des Weltfrieden hat niemand der IS zu überleben! So wäre das zu händeln! (engl. In the interest of world peace, no one should survive the IS! That is how it should be handled!)

Further cynical or sarcastic tweets make correct classification difficult without general knowledge, such as the following tweet in the criticism category of the second subtask.

> DDR 2.0 läßt grüßen (engl. Greetings from DDR 2.0)

Moreover, the following tweet, which is assigned to the violence category of subtask 3, is difficult to classify without further context.

> Spende Steakmesserset.... (engl. Donate a steak knife set...)

Very short posts also pose a problem when, for example, calls to action consist of only a few words, e.g. *schießen !!!* (engl. *shoot!!!*) and *einfach nur drüberfahren* (engl. *just run them over*).

7 Conclusion

This paper presents the pilot edition of the GermEval shared task for harmful content detection. This shared task comprised three subtasks: the detection of (1) calls to action, (2) various attacks on the DBO, and (3) worrying violence-related statements. For this purpose, new German-language Twitter datasets (approx. 9k–11k tweets) were constructed, most of which originate from the environment of a right-wing extremist group. A total of eleven teams participated, mainly using German-language or multilingual BERT models. The best systems achieved macro-average F_1-scores of up to 87% in subtasks 1 and 3 and up to 71% in subtask 2. German BERT variants, ensembles of various BERT-based models and modern open-source LLMs developed specifically for German, some of which were fine-tuned using training data supplemented with synthetic examples, were particularly successful.

Despite promising results, there is still room for improvement. For example, multi-task learning was not tested by any team, although it has already shown success for related classification problems (e.g. Kancharla et al., 2025). In addition, the datasets have limitations in terms of low inter-annotator agreement and a substantial underrepresentation of harmful content. For a possible second edition of the shared task, the annotation of further data is therefore planned, which should also enable the third task to be aligned as a fine-grained classification in the sense of a finer differentiation of different types of violent statements.

References

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

Issa Alsmadi and Keng Hoon Gan. 2019. Review of short-text classification. *International Journal of Web Information Systems*, 15(2):155–182.

Gabriel Assis, Annie Amorim, Jonnathan Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. 2024. Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models? In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 1, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen Technical Report. *arXiv preprint*.

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, An-

drea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, Gullal S. Cheema, Dilshod Azizov, and Preslav Nakov. 2023. The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, volume 13982, pages 506–517. Springer Nature Switzerland, Cham.

Bundesamt für Verfassungsschutz. 2024. Rechtsextremismus im Internet: Gefahren digitaler Agitation und Radikalisierung. Retrieved 29 August 2025.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 109–117, Nagoya, Japan. Asian Federation of Natural Language Processing.

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.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27.

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

Paulina Garcia Corral, Avishai Green, Hendrik Meyer, Anke Stoll, Xiaoyue Yan, and Myrthe Reuver. 2024. A Few Hypocrites: Few-Shot Learning and Subtype Definitions for Detecting Hypocrisy Accusations in Online Climate Change Debates. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and Short Papers*, pages 45–60, Vienna, Austria. Association for Computational Linguistics.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, pages arXiv–2407.

Duden. 2025. Subversion. Retrieved 29 August 2025.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. ModernGBERT: German-only 1B Encoder Model Trained from Scratch. *arXiv preprint*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Mayur Gaikwad, Swati Ahirrao, Ketan Kotecha, and Ajith Abraham. 2022. Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques. *IEEE access : practical innovations, open solutions*, 10:104829–104843.

Stephanie Gross, Johann Petrak, Louisa Venhoff, and Brigitte Krenn. 2024. GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora. In *Proceedings of GermEval 2024 Task 1 GerMS-detect Workshop on Sexism Detection in German Online News Fora (GerMS-detect 2024)*, pages 1–9, Vienna, Austria. Association for Computational Lingustics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *Preprint*, arXiv:2006.03654.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of Experts. *arXiv preprint*.

Bharath Kancharla, Prabhjot Singh, Lohith Bhagavan Kancharla, Yashita Chama, and Raksha Sharma. 2025. Identifying aggression and offensive language in code-mixed tweets: A multi-task transfer learning approach. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 122–128, Abu Dhabi. Association for Computational Linguistics.

Muhammad Shahid Khan, Muhammad Shahid Iqbal Malik, and Aamer Nadeem. 2024. Detection of violence incitation expressions in Urdu tweets using convolutional neural network. *Expert Systems with Applications*, 245:123174.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Damir Korenčić, Berta Chulvi, X Bonet Casals, Mariona Taulé, Paolo Rosso, and Francisco Rangel. 2024. Overview of the oppositional thinking analysis pan task at clef 2024. In *Working Notes of CLEF*, pages 2462–2485, Grenoble, France. CLEF.

Vincent Kums, Florian Meyer, Luisa Pivit, Uliana Vedenina, Jonas Wortmann, Melanie Siegel, and Dirk Labudde. 2025. A novel dataset for classifying German hate speech comments with criminal relevance. In *Proceedings of the the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 41–52, Vienna, Austria. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

Jens Lemmens, Jens Van Nooten, Tim Kreutz, and Walter Daelemans. 2022. CoNTACT: A Dutch COVID-19 adapted BERT for vaccine hesitancy and argumentation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6837–6845, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Steven Loria. 2018. Textblob Documentation. Retrieved 29 August 2025.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, pages 29–32, Hyderabad India. ACM.

Oxford University Press. 2024. Call to action. Retrieved 29 August 2025.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE access : practical innovations, open solutions*, 11:30575–30590.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to Action on Social Media: Potential for Censorship and Social Impact. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 36–44, Stroudsburg, PA. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. BLP-2023 task 1: Violence inciting text detection (VITD). In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 255–265, Singapore. Association for Computational Linguistics.

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. 2023. Zero and Few-Shot Hate Speech Detection in Social Media Messages Related to Earthquake Disaster. In *Proceedings of the 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, Istanbul, Turkiye. IEEE.

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2022. Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–7, Kolkata India. ACM.

Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 shared task on statement segmentation in German easy language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)*, pages 1–14, Vienna, Austria. Association for Computational Lingustics.

Wassiliki Siskou, Clara Giralt Mirón, Sarah Molina-Raith, and Miriam Butt. 2022. Automatized detection and annotation for calls to action in Latin-American social media postings. In *Proceedings of*

the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 65–69, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

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

Katja Thorwarth. 2023. „Wir schießen den Weg frei“ - bereitet die AfD-Sprache den Boden für rechten Terror mit? *Frankfurter Rundschau*. Retrieved 29 August 2025.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–8, Vienna, Austria. German Society for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*.

A Appendix

A.1 Details about class distribution after data augmentation

The dataset used in the shared task for the second subtask, the detection of attacks on the DBO, consisted of tweets from the German Twitter network of a right-wing extremist group as well as additional tweets from the datasets by Demus et al. (2022) and Kums et al. (2025). The data augmentation led to an increase in the number of tweets in all four categories. The number of tweets before and after data enrichment for the second subtask can be found in Table 9. The data for the third subtask, the detection of problematic, violence-related statements, was only enriched with data by Kums et al. (2025), which increased the number of tweets in the TRUE category.

A.2 Details about the data format

An example of a row from the training data set for the first subtask, the detection of calls to action, can be found in Table 10.

A.3 Prompt engineering for the baseline of subtask 3

For classification in subtask 3, a system prompt and a user prompt were fed into the model Qwen2.5. The system prompt consisted of a role assignment, a task description, a definition of worrying violence-related statements, and a description of the desired output:

System Prompt

Instruction: You are a filter system in a social network that specialises in monitoring German tweets. Your task is to detect worrying positive attitudes towards violence. Consider the following definitions: A tweet contains a worrying violence-related statement if it trivialises, advocates, glorifies, incites violence or announces the use of violence. Violence is understood to mean active, aggressive behaviour, i.e. the infliction of physical, sexual or psychological violence. If you predict a violence-related statement, your response is “true”. If you predict a no-violence-related statement, you respond with “false”. You only respond with one single word and not more!

The user prompt was composed of the following components: repetition of the task description, repetition of the desired output, examples of the classification task for in-context learning and the tweet to be classified.

User Prompt

Determine whether the input tweet contains a worrying violence-related statement. Respond with a single word. Do not provide explanations or notes. Your only possible responses, regarding if the provided text contains a worrying violence-related statement are true and false. Examples of classification are: <Examples> Input: <Tweet>

The following six tweets, which had been presented to the annotators, were provided as examples:

- @UserXY Wenn ich diesen Typ schon sehe, könnte ich ihn an die Wand stellen response: “true”
- Knallt das ganze linksgrün Dumme Antifa

Table 9: Class distribution in the total data set for subtasks 2 and 3 before and after data augmentation.

Subtask	categories	Newly collected tweets	After data augmentation
Subtask 2: Attacks on DBO	subversive	29 (0.31)	85 (0.80)
	agitation	48 (0.52)	447 (4.2)
	criticism	1072 (11.52)	1149 (10.79)
	nothing	8158 (87.65)	8967 (84.21)
	total	9307 (100)	10648 (100)
Subtask 3: Violence Detection	true	620 (5.67)	805 (7.24)
	false	10313 (94.339)	10313 (92.76)
	total	10933 (100)	11118 (100)

Table 10: Example row from the training data set for the first subtask, the detection of calls to action.

id	description	C2A
1096188353752920	Einfach überfahren wo ist das Problem ?	TRUE

Gesindel einfach ab! #LinksfaschistenStoppen *response: "true"*

- Man stelle sich vor, eine Gruppe on 40 jungen MÄNNERN [. . .], teilweise BEWAFFNET [. . .], versucht ILLEGAL eine Grenze zu übertreten, [. . .] und dann wundert man sich, wenn das Feuer gerechtfertigterweise eröffnet wird. *response: "true"*
- @UserXY endlich hat jemand den Mut ergriffen und dem Mistkerl die Fresse poliert. Der hat alles schlimme der Welt und mehr verdient dieser erbärmliche Perverse und geistig Behinderte #Volksverräter *response: "true"*
- @UserXY DU BIST TEIL dieser ganz GEZIELT INSZENIERTEN LÜGE über Lynchjustiz und Aufruf zu Mord und versuchst alle durch FREI ERFUNDENE HORRORGeschichten in PANIK zu VERSETZEN! #Lügenpresse *response: "true"*
- [. . .] Der 27- jährige Syrer wird verdächtigt, zwischen 2011 und 2014 als Mitglied von Jabhat al- Nusra bzw. Kataib Al Faroug im syrischen Bürgerkrieg gegen Anhänger bzw. Truppen des Assad- Regimes gekämpft und auf diese geschossen zu haben, sagte Peter Oehm, der Leiter des Landesamtes für Verfassungsschutz. *response: "false"*