NLPeace@GermEval Shared Task 2025: fine-tuned BERT vs. prompted LLMs for German hate speech detection

Patrick Göttfert, Raphael Huber, Fabio Mariani

Angaben zur Veröffentlichung / Publication details:

Göttfert, Patrick, Raphael Huber, and Fabio Mariani. 2025. "NLPeace@GermEval Shared Task 2025: fine-tuned BERT vs. prompted LLMs for German hate speech detection." In *KONVENS: 21th Conference on Natural Language Processing (KONVENS 2025); proceedings of the conference - volume 2: workshops, Hildesheim, Germany, 9.-12. September 2025*, edited by Christian Wartena and Ulrich Heid, 350–56. Hannover: HsH Applied Academics. https://doi.org/10.25968/opus-3679.

NLPeace@GermEval Shared Task 2025: Fine-Tuned BERT vs. Prompted LLMs for German Hate Speech Detection

Patrick Göttfert
University of Augsburg
patrick.goettfert@uni-a.de

Raphael Huber
University of Augsburg
raphael.huber@uni-a.de

Fabio Mariani
University of Augsburg
fabio.mariani@uni-a.de

Abstract

We present our participation in the GermEval 2025 Shared Task on Harmful Content Detection in German social media, focusing on binary classification of calls to action (C2A) and violent content (VIO). We compare fine-tuned BERT models with prompting-based approaches using large language models (Mixtral, Qwen) in zero-shot and few-shot settings. Our results show that supervised BERT classifiers clearly outperform prompting methods, reaching macro-F1 scores up to 0.78 on the test set. While LLM prompting was competitive in some cases, it suffered from inconsistent outputs and high sensitivity to prompt wording. Overall, our findings highlight the reliability of fine-tuning for this task and the need for improved prompt design and strict separation between training and trial data.

1 Introduction

The proliferation of harmful content on social media, including hate speech, extremist propaganda, and incitement to violence, poses significant societal and technological challenges. Prior research shows that abusive language is frequent on platforms such as **X** (formerly Twitter) and can cause psychological harm and escalate to real-world violence (Waseem and Hovy, 2016). While hate speech is an important subset, other forms—such as calls to action and attacks on democratic institutions—are equally critical to detect and moderate.

To advance research in this area, the **Harmful Content Detection in Social Media** shared task[1] was organized on Codabench, comprising three subtasks evaluated via macro-F1 score (Felser et al., 2025). We focused on the two binary classification subtasks: **Call to Action (C2A)** and **Violence Detection (VIO)**. In C2A, the goal is to detect statements encouraging specific behavior. For example:

> "Jeder bringt zur morgigen Demonstration mindestens 10 Leute mit, denn eine Bürgerbewegung ist nur so stark wie ihre Anhänger!!!"

The VIO subtask targets tweets expressing a disturbingly positive attitude towards violence:

> "Knallt das ganze linksgrün dumme Antifa-Gesindel einfach ab! #LinksfaschistenStoppen"

Both subtasks are framed as binary classification problems, labeling each tweet as `True` (harmful) or `False` (not harmful). A newly constructed dataset and baseline models were provided for evaluation.

In this work, we explore these tasks by fine-tuning BERT (Devlin et al., 2019) models and experimenting with prompting techniques using state-of-the-art large language models (LLMs). Our main contribution is a comparative evaluation of both approaches on the shared task dataset.

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 describes our methodology, Section 4 reports results, and Section 5 concludes with directions for future work.

2 Related Work

Li et al. (2024) evaluate ChatGPT's performance in annotating hateful, offensive, and toxic comments. They report high agreement with human annotators on non-harmful texts but substantially lower consistency for harmful content. Moreover, the study highlights the model's strong prompt sensitivity and tendency to produce extreme probability estimates. These findings underscore the importance of careful prompt engineering and class design, an aspect we also encountered when prompting LLMs for binary classification.

Alatawi et al. (2023) show that a BiLSTM classifier trained on embeddings derived from a large

[1] https://www.codabench.org/competitions/4963/#/pages-tab

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 350–356, 2025

hate speech corpus achieves strong results, while fine-tuning BERT further increases F1-scores to 96%. However, the authors note that BERT struggles with intentionally misspelled or coded hate terms, underlining the need for domain-adapted representations.

In addition, Fillies et al. (2025) present the first German-language TikTok dataset focusing on far-right extremist discourse. Their dataset comprises over 10,000 comments annotated with a fine-grained taxonomy of 32 labels, capturing diverse targets, intensities, and types of hate speech. This work highlights the importance of platform-specific resources and the challenges of moderating content in dynamic environments like TikTok, which shows notable differences compared to the German Twitter data we used.

Beyond text-only approaches, multimodal research addresses the growing challenge of harmful content embedded in images and memes. Arya et al. (2024) propose a detection method based on Contrastive Language-Image Pre-Training (CLIP) combined with prompt engineering. Their model achieves an accuracy of 87.42%, demonstrating the effectiveness of joint vision-language reasoning. While our work focuses on textual classification, such multimodal strategies are likely to become increasingly important as harmful content diversifies in form and medium.

Furthermore Piot et al. (2024) present MetaHate, a unified hate speech dataset combining 36 existing corpora into a harmonized binary classification scheme with over 1.2 million labeled instances. Their benchmarks show that transformer-based models, particularly BERT, achieve the strongest performance (Macro-F1 ≈ 0.80), underlining the value of large, diverse datasets for robust and generalizable hate speech detection.

Moreover, Awal et al. (2021) propose AngryBERT, a multitask learning model for hate speech detection. It jointly learns the primary task of hate speech classification alongside two related secondary tasks: emotion classification and target identification. Using BERT as a shared layer, the model leverages features learned across tasks to improve overall performance. AngryBERT is evaluated on three public hate speech datasets (WZ-LS, DT, and FOUNTA), achieving higher F1-scores than both single-task BERT models and other multitask baselines. These results highlight the benefit of incorporating sentiment and target information for more robust hate speech detection. While our approach focuses on single-task BERT fine-tuning for C2A and VIO, the multitask strategy explored in AngryBERT suggests a promising direction for future work.

3 Methodology

In this section, we describe the dataset and models used in our approach to the shared task, including the experimental setup for each subtask.

3.1 Data

All subtasks rely on a newly annotated dataset derived from the German Twitter network of an extremist group, collected between 2014 and 2016.[2] The corpus comprises approximately 11,500 tweets annotated by trained coders with expertise in harmful content. To ensure privacy, all mentions of other users were anonymized.

The data is split into training, trial, and test sets. The test set labels were withheld for final evaluation. Each entry contains an ID, the tweet text, and a binary label indicating whether the tweet is harmful.

For subtask C2A, the training set contains 6,840 tweets and the trial set ~1,000. For VIO, there are 7,783 training tweets and ~1,000 trial examples. The trial set was used during development to compare and fine-tune approaches, while the test set served exclusively for final submission. Crucially, we discovered only after the submission deadline that the trial set overlapped with the training data, which likely led to inflated evaluation scores during development.

3.2 Subtask 1: C2A

We evaluated both traditional machine learning classifiers and transformer-based models for the C2A subtask.

Baseline Classifiers As reference systems, we adapted the baseline provided for the C2A subtask, which originally employed a Gradient Boosting Classifier for evaluation on the trial set. We further experimented with Logistic Regression, Gaussian Naive Bayes, and K-Nearest Neighbors, all trained on sentence embeddings generated by the `distiluse-base-multilingual-cased-v2` encoder.[3] This model is a distilled version of

[2] https://github.com/Communication-Forensics-Lab/harmful-content-detection

[3] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

the multilingual Universal Sentence Encoder, optimized for producing sentence-level embeddings. Preprocessing included lemmatization and tokenization with SpaCy[4] and polarity feature extraction using TextBlobDE[5]. This variant was not fine-tuned.

BERT-based Classifiers We fine-tuned two BERT variants:

- **`bert-base-uncased`**[6]: implemented in PyTorch as a sequence classifier with a linear layer and sigmoid activation. Trained for 3 epochs with a batch size of 8 and a learning rate of 1e-5, validating on the trial set.
- **`bert-base-german-cased`**[7]: a cased German BERT pretrained on large German text corpora, fine-tuned using the Hugging Face `transformers` library, trained for 3 epochs (learning rate 2e-5, batch size 8, AdamW optimizer), also validating on the trial set.

LLM-based Classifiers We tested `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024)[8], an instruction-following LLM capable of processing German prompts without additional task-specific fine-tuning, in zero-shot and few-shot prompting (4 examples), following its instruction format, with inference via the Hugging Face pipeline (16-bit quantization). Outputs were post-processed to extract labels.

We also evaluated `Qwen2.5-32B-Instruct` (Bai et al., 2023)[9], a large multilingual instruction-following LLM supporting German input, with the same zero-shot prompt, including a variant with a minor spelling error. Qwen allows separate system and user prompts, which we leveraged to structure inputs.

3.3 Subtask 2: VIO

Due to structural similarity, we reused the C2A scripts with minimal changes for VIO labels. For LLMs, we introduced a dedicated prompt (see Appendix A.2) and evaluated both zero-shot and few-shot settings.

We also explored zero-shot chain-of-thought prompting with Qwen by adding the instruction:

```
The last TRUE or FALSE should be
the answer. Let's think step by
step.
```

This prompted step-by-step reasoning. Due to longer processing time, we reduced the batch size to 1. Additionally, we evaluated the official baseline prompt using our Qwen inference script.

4 Results

Table 1 presents the Macro-F1 scores for both subtasks —C2A and VIO — evaluated on the trial set. We compare several classification techniques, including baselines, transformer-based models, and LLM prompting methods.

For both subtasks, the provided official baseline based on a Gradient Boosting Classifier serves as a strong starting point, achieving F1-scores of 0.65 (C2A) and 0.59 (VIO). However, when experimenting with alternative classifiers using the same sentence embeddings, we observed lower performance across the board. Logistic Regression yielded the strongest results among these alternatives, with scores of 0.61 for C2A and 0.55 for VIO.

Our BERT-based models substantially outperformed the baselines. The best-performing model on C2A was the fine-tuned `bert-base-uncased` variant, achieving an F1-score of 0.92. For the VIO subtask, the fine-tuned `bert-base-german-cased` produced the highest score of 0.98, marking the best result overall across all configurations.

Prompt-based classification using LLMs showed mixed results. The `Mixtral-8x7B-Instruct` model, both in zero-shot and few-shot setups, achieved F1-scores around 0.52–0.53 for C2A and 0.70 for VIO. The `Qwen2.5-32B-Instruct` model generally performed slightly lower on C2A (with scores ranging from 0.46 to 0.52), but matched or surpassed Mixtral on VIO, reaching up to 0.76 using the prompt from the official baseline.

Furthermore, correcting a minor spelling error in the prompt improved the F1 score by 0.06, as seen in the table, showing that even small prompt changes can significantly affect LLM performance.

Overall, transformer-based fine-tuned models clearly outperform prompt-based approaches in this setting, especially for the C2A task. Nonetheless,

[4] https://spacy.io/
[5] https://textblob.readthedocs.io/
[6] https://huggingface.co/google-bert/bert-base-uncased
[7] https://huggingface.co/google-bert/bert-base-german-cased
[8] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[9] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

Technique	C2A	VIO
Baseline from C2A (using Gradient Boosting Classifier)	0.65	0.59
Adapted Baseline using K-Neighbours	0.55	0.39
Adapted Baseline using Gaussian Naive Bayes	0.53	0.50
Adapted Baseline using Logistic Regression	0.61	0.55
Fine-tuned bert-base-uncased	**0.92**	0.71
Fine-tuned bert-base-german-cased	0.85	**0.98**
Mixtral	0.52	0.70
Mixtral with few-shot	0.53	0.70
Qwen with faulty spelling	0.46	-
Qwen with corrected spelling	0.52	0.69
Qwen with prompt from baseline	-	0.76
Qwen with zero-shot chain-of-thought	-	0.70

Table 1: All Macro-F1 scores evaluated on the Trial Set (rounded to two decimal places). For each subtask and model category, the best score is underlined, while the overall best score per subtask is highlighted in **bold**.

Technique	C2A	VIO
Fine-tuned bert-base-german-cased	**0.78**	**0.76**
Mixtral with few-shot	0.54	0.71

Table 2: All F1-scores were evaluated on the test set. The overall best score per subtask is highlighted in **bold**.

few-shot prompting remains a promising alternative for zero-resource or multilingual transfer scenarios.

Finally, Table 2 summarizes the results obtained on the official test set using our final submitted systems. As expected, our fine-tuned BERT model consistently achieved the highest F1-scores across both subtasks, with 0.78 for C2A and 0.76 for VIO. In contrast, the best-performing LLM-based method (Mixtral in few-shot mode) yielded competitive but lower scores, particularly in the VIO subtask with 0.71.

We selected these two systems for final submission to Codabench in order to directly compare transformer-based fine-tuning with prompting-based LLM approaches. Among all our BERT-based experiments, the custom script consistently delivered the strongest results, while among LLM-based runs, the few-shot Mixtral model performed best—except for a single run using the baseline prompt provided in the task documentation. However, we opted not to use that baseline prompt in our final submission, as we aimed to showcase the effectiveness of our own prompt design and modeling choices.

After our experiments, we discovered that the trial set was actually a subset of the training set, a fact that became apparent only after the experimentation phase. This overlap introduces a risk of overfitting when evaluating models on the trial set, as the models may have already seen this data during training. Consequently, the trial set scores likely overestimate real-world performance, and the lower scores on the unseen test set are consistent with this expectation.

Additionally, we observed that prompting-based approaches using LLMs did not always follow our explicit instruction to return only the binary labels "TRUE" or "FALSE." Despite phrasing the prompts clearly (e.g., "Respond only with 'TRUE' or 'FALSE'. No explanation."), the models occasionally produced additional output such as punctuation (e.g., ", TRUE") or full sentences requiring further parsing to extract only the labels.

5 Conclusion and Future Work

In this work, we participated in the GermEval 2025 Shared Task on Harmful Content Detection, focusing on binary classification subtasks C2A and VIO. We compared fine-tuned transformer models and prompting-based approaches with LLMs.

Our experiments showed that transformer models, particularly our fine-tuned BERT, outperformed prompting methods in classification accuracy. While LLMs like Mixtral and Qwen delivered competitive results in few-shot and chain-of-

thought configurations, they remained less reliable overall. However, LLMs offered advantages in development speed and required no additional training, albeit with challenges such as inconsistent output formatting.

Overall, our findings suggest that for small- to medium-sized datasets, supervised fine-tuning remains the most effective approach. Prompting with LLMs requires further refinement to improve reliability.

In the following, we outline directions for future work and improvements based on our experimental findings. For prompting methods, more systematic prompt engineering such as testing different templates and incorporating domain-relevant few-shot examples is needed to improve reliability.

On the supervised side, future work could explore alternative transformer architectures like RoBERTa (Liu et al., 2019) or DeBERTa (He et al., 2021). Addressing class imbalance and refining loss functions may reduce performance variability.

Ensuring clear separation of training, development, and trial data is crucial for fair evaluation. Future error analysis is needed to better understand the notably different performance of the two BERT models, particularly whether the variation stems from differences in language coverage, sensitivity to domain-specific patterns, or divergent handling of implicit harmful language.

References

Hind Saleh Alatawi, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using BERT and hate speech word embedding with deep model. *Appl. Artif. Intell.*, 37(1).

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M. Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12:22359–22375.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angrybert: Joint learning target and emotion for hate speech detection. *Preprint*, arXiv:2103.11800.

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. *Preprint*, arXiv:2309.16609.

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. Association for Computational Linguistics.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the germeval 2025 shared task on harmful content detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany.

Jan Fillies, Esther Theisen, Michael Hoffmann, Robert Jung, Elena Jung, Nele Fischer, and Adrian Paschke. 2025. A novel german tiktok hate speech dataset: far-right comments against politicians, women, and others. *Discover Data*, 3.

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. *CoRR*, abs/2401.04088.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans. Web*, 18(2):30:1–30:36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:2025–2039.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

A Appendix

This appendix provides supplementary materials, including our changes after the submission deadline and our prompt templates.

A.1 Experiments Conducted After the Codabench Phase

After the official Codabench phase ended, we conducted a few additional experiments. We added a few-shot script and corresponding result for the C2A task. In this context, the result for the baseline with the gradient boosting classifier and the results for the K-Nearest Neighbors classifier was updated due to a correction.

A.2 Prompt Templates

A.2.1 C2A: Zero-shot Prompt Mixtral

```
<s>[INST] You are a classification model. The input text is in German.
Output only 'TRUE' if it contains hate speech, otherwise 'FALSE'.
Respond only with 'TRUE' or 'FALSE'. Nothing else! [/INST]
```

A.2.2 C2A: Zero-shot Prompt Qwen with correct spelling

```
 [{"role": "system", "content": "You are a classification model."},
{"role": "user", "content": f"The input text is in German. Output only 'TRUE' if it
    contains hate speech, otherwise 'FALSE'. Respond only with 'TRUE' or 'FALSE'.
    Nothing else!{text}"}]
```

A.2.3 C2A: Zero-shot Prompt Qwen with faulty spelling

```
[{"role": "system", "content": "You are a classification model."},
{"role": "user", "content": f"The input text is in German. Output only 'TRUE' if it
    contains hate speech, otherwise 'FARLSE'. Respond only with 'TRUE' or 'FALSE'.
    Nothing else!{text}"}]
```

A.2.4 C2A: Few-shot Prompt Mixtral

```
[INST] You are a classification model. The input text is in German.
Output only 'TRUE' or 'FALSE'. No explanation. If the following text contains hate
    speech, output 'TRUE'.
Otherwise output 'FALSE'.

Text: die aus sicheren herkunftsländern nach hause schicken.
Label: TRUE
Text: Tja, wohl etwas schwierig, jetzt das Geld per Chip in die Heimat zu überweisen
Label: FALSE
Text: vergrabt ein totes Schwein, hat in Spanien auch den Bau der Moschee verhindert
    .
Label: TRUE
Text: In der Regierung haben doch schon alle das sonnenkönigsyndrom, lol.
Label: FALSE
```

A.2.5 VIO: Zero-shot Prompt Mixtral

```
[INST] You are a classification model. The input text is in German.
Output only 'TRUE' if it expresses a disturbingly positive attitude towards violence
(e.g., support, incitement, glorification, or willingness to commit violence).
Otherwise output 'FALSE'. Respond only with 'TRUE' or 'FALSE'. No explanation. [/
    INST]
```

A.2.6 VIO: Few-shot Prompt Mixtral

```
[INST] You are a classification model. The input text is in German.
Output only 'TRUE' or 'FALSE'. No explanation. If the following text expresses a
    disturbingly positive
attitude towards violence (e.g. glorification, incitement, support, or personal
    willingness), output 'TRUE'.
Otherwise, output 'FALSE'.
```

```
Text: schau Dir die Kinder an. Die werden benutzt! Die sollte man rausholen, aber
    die Eltern dort lassen
Label: TRUE
Text: Karl Valentin hat es auf den Punkt gebracht: "... und das Ritterfräulein
    Annerl ward von einem Neger schwanger. Doch der Ritter sagt nur: Scheiß,
    Hauptsach is, es wird ka Preiß!"
Label: FALSE
Text: Oh Graus!!! Vom Gehöft Sollte man Diese Jenigen jagen!!!??????
Label: TRUE
Text: Yeah Überraschungen - verkündet ihr endlich die Abtrennung von Deutschland?
Label: FALSE
```

A.2.7 VIO: Zero-shot Prompt Qwen

```
[{"role": "system", "content": "You are a classification model."},
{"role": "user", "content": f"The input text is in German. Output only 'TRUE' if it
    expresses a disturbingly positive attitude towards violence (e.g., support,
    incitement, glorification, or willingness to commit violence). Otherwise output
    'FALSE'. Respond only with 'TRUE' or 'FALSE'. No explanation. {text}"}]
```

A.2.8 VIO: Zero-shot chain-of-thought Prompt Qwen

```
[{"role": "system", "content": "You are a classification model."},
{"role": "user", "content": f"The input text is in German. Output only 'TRUE' if it
    expresses a disturbingly positive attitude towards violence (e.g., support,
    incitement, glorification, or willingness to commit violence). Otherwise output
    'FALSE'. The last TRUE or FALSE should be the answer. Let's think step by step.{
    text}"}]
```

A.2.9 VIO: Zero-shot with baseline prompt

```
[{"role": "system", "content": "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.\nConsider 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!"},
{"role": "user", "content": f"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. \
    nExamples of classification are:\n@UserXY Wenn ich diesen Typ schon sehe, könnte
     ich ihn an die Wand stellen response: \"true\"\nKnallt das ganze linksgrün
    Dumme Antifa Gesindel einfach ab! #LinksfaschistenStoppen  response: \"true\"\
    nMan 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\"\n@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\"\n@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! #Lü
    genpresse  response: \"true\"\n\".[.] 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\" Input: {text}"}]
```

A.3 Parameter of the Baseline-Classifier

All classifiers were used with the default parameters of the Python library `scikit-learn`. For `LogisticRegression`, we set `max_iter` to 10000 and `random_state` to 42.