TUM NLP Group at GermEval 2025 Shared Task on Candy Speech Detection: Small Dose, Big Effect: Leveraging Synthetic Data for Candy Speech Detection

Sophie Luise Opheiden, Elisei Shushpanov, Faeze Ghorbanpour
Department of Informatics
Technical University of Munich
{sophie.opheiden, e.shushpanov, faeze.ghorbanpour}@tum.de

Abstract

This paper presents our submission to the GermEval 2025 Shared Task on Candy Speech Detection. We fine-tune a multilingual transformer model (mDeBERTa-v3) on original and LLM-augmented data using prompt-based dual-class generation. Our best model, trained with 10% synthetic data per class, achieved an F1-score of 0.8785 on the test set, ranking 6th out of 20 submissions. Further augmentation did not improve performance, underlining the need for carefully targeted augmentation.

1 Introduction

Over the past two decades, social media platforms have risen to be one of the main communication channels across generations. With the increasing number of users, the need to monitor online discourse has grown — particularly to prevent the misuse of anonymity that these platforms provide. Consequently, numerous methods for detecting and censoring negative speech (e.g., hate speech or harmful language) on social media have been developed, allowing social media platform providers to enforce community guidelines.

In contrast, less attention has been given to identifying and promoting positive and supportive discourse in online communities (Clausen et al., 2025). Yet, this research field holds relevance for practical applications and academia: Candy speech detection complements negative speech detection by actively identifying and promoting positive communication. This can help foster a more welcoming and empathetic online culture, especially on platforms like YouTube, where comments can significantly impact both creators and viewers.

Sentiment detection in social media is complex: The nature of social media content, such as YouTube video comments, is informal and may intentionally deviate from standard orthography to enhance expressiveness and convey hyperbolic sentiment beyond the literal meaning of the word. This makes processing of the input text data not trivial, potentially negatively impacting downstream model performance.

The GermEval Shared Task 2025 on Candy Speech Detection (Clausen et al., 2025) invited researchers to detect candy speech in YouTube comments. We participated in Subtask 1: Coarse-Grained Classification, which targets the identification of *"Flausch"* speech in German YouTube comments as a whole. Our best approach fine-tuned a transformer-based model on a dataset augmented by 10%. Notably, our experiments showed that increasing the amount of synthetic data beyond this point did not lead to further improvements.

2 Dataset

The dataset consists of German-language YouTube comments. The source videos vary widely in content, style, audience, and creators, resulting in diverse topic distributions. Notably, the final test set includes previously unseen topics, highlighting the importance of robust model generalization.

The dataset is imbalanced, with only 29% of the comments labeled as *"Flausch"* and 71% as *"not Flausch"*. It also contains highly informal language, including intentional spelling deviations (e.g., "suuuuper") to express emphasis or hyperbole. Most comments are short: Fifty percent contain six words or fewer and 32 characters or fewer.

3 Methodology

This section introduces a formal problem definition of the GermEval Challenge Task 1 (Coarse-Grained Classification). Subsequently, we discuss experiment design and utilized methods with justifications for the steps taken in the experiment.

3.1 Problem Definition

Candy speech, also known by its German name *"Flausch"*, has been formally defined by the

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 438–443, 2025

GermEval 2025 organizing committee as an "expression of positive attitudes on social media toward individuals or their output (videos, comments, etc.)" (Clausen et al., 2025).

Subtask 1 (Coarse-Grained Classification) is framed as a binary classification problem: given a YouTube comment $c \in C$, the goal is to learn a function

$$f : C \to \{0, 1\}$$

that predicts whether c contains candy speech ($f(c) = 1$) or not ($f(c) = 0$).

A labeled training set is provided in the form:

$$D_{\text{train}} = \{(c_i, y_i)\}_{i=1}^{N}, \quad y_i \in \{0, 1\} \qquad (1)$$

The classifier f is trained on D_{train} and evaluated on an unseen test set $D_{\text{test}} = \{(c_j, y_j)\}_{j=1}^{M}$.

3.2 Model Selection

We conducted an exploratory comparison of popular transformer models to identify the strongest baseline. The models include: mBERT[1] (Devlin et al., 2018), a multilingual bidirectional encoder representation from transformers; XLM-RoBERTa[2] (Conneau et al., 2019), a robustly optimized BERT pretraining model; and mDeBERTaV3[3], a multilingual version of DeBERTaV3(He et al., 2021a) which pretrain DeBERTa(He et al., 2021b) using gradient-disentangled embedding sharing. DeBERTa itself is a decoding-enhanced BERT architecture with disentangled attention.

Grid search and Optuna[4] (Akiba et al., 2019) were utilized to find the best hyperparameters. Table 1 compares the performance of all models across various hyperparameter settings. We used macro-averaged metrics for the exploration, as this allowed us to assess the performance across all classes, regardless of class imbalance. Choosing the best model, we employed mDeBERTaV3, a transformer-based encoder architecture developed by Microsoft.

mDeBERTaV3 is pretrained across 100+ languages. Known for its strong performance in multilingual natural language processing (NLP) tasks, it is able to capture fine-grained syntactic and semantic patterns—an important property for detecting certain sentiment in stylistically distorted social media comments.

Our final baseline model was fine-tuned with a learning rate of `3e-5`, batch size `16`, `3` epochs, and weight decay `0.01`. All textual inputs were tokenized with the model's associated tokenizer.

3.3 Data Augmentation

Data augmentation is used to artificially expand a training dataset by creating modified versions of existing data. Using this technique, generalization to unseen data may be improved as the size of the training set is increased. Three methods were implemented:

Synonym Replacement is a lexical data augmentation technique, in which randomly selected words in a sentence are replaced with semantically similar alternatives using word embedding models.

Back translation involves translating a sentence into an intermediate language (e.g., English) and then translating it back to the original language. It is particularly useful for creating diverse training examples that differ in structure and phrasing but convey the same semantic content.

Large Language Model (LLM) based Dual-Class Augmentation is a prompt-based generative approach. In contrast to single-class augmentation, it includes examples from both classes in the prompt, allowing the model to better learn the decision boundary between classes (Ibrahim et al., 2025). Each prompt combined five random examples from each class, and the model was instructed to generate a new German comment for a given label (Table 2). We introduced the classification labels and their definitions in the prompt to provide clearer context for the generation task and explicitly associate the label *"Flausch"* with its meaning. This approach was intended to help the model better understand the task requirements while also reducing prompt length, as it eliminated the need to repeatedly replace *"Flausch"* with its full definition in each instruction. We augmented both *"Flausch"* and *"not Flausch"* classes equally to preserve the original class distribution.

To study the effect of class imbalance on model performance, we created training sets with different *"not Flausch"*/*"Flausch"* ratios (70/30, 65/35, 60/40, 50/50) utilizing synonym replacement and back translation. Synonym replacement was implemented leveraging a pre-trained

[1] https://huggingface.co/google-bert/bert-base-multilingual-cased

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

[3] https://huggingface.co/microsoft/mdeberta-v3-base

[4] https://github.com/optuna/optuna

Model	Method	Accuracy	F1	Precision	Recall	Hyperparameters
mBERT	Grid Search	0.8779	0.8392	0.8714	0.8093	LR=2e-5, BS=16, Epochs=3, WD=0.01
mBERT	Grid Search	0.8781	0.8410	0.8504	0.8318	LR=3e-5, BS=16, Epochs=3, WD=0.01
mBERT	Grid Search	0.7910	0.7635	0.6798	0.8707	LR=5e-5, BS=8, Epochs=4, WD=0.01
mBERT	Grid Search	0.8763	0.8411	0.8370	0.8453	LR=1e-5, BS=32, Epochs=5, WD=0.01
mBERT	Optuna	0.8308	0.8385	0.8589	0.8308	LR=4.28e-5, BS=8, Epochs=5, WD=0.0706, WR=0.167
mBERT	Optuna	0.8171	0.8275	0.8624	0.8171	LR=4.35e-5, BS=16, Epochs=2, WD=0.0551, WR=0.174
XLM-RoBERTa	Grid Search	0.8790	0.8445	0.8414	0.8477	LR=2e-5, BS=16, Epochs=3, WD=0.01
XLM-RoBERTa	Grid Search	0.8789	0.8431	0.8467	0.8396	LR=3e-5, BS=16, Epochs=3, WD=0.01
XLM-RoBERTa	Grid Search	0.8844	0.8513	0.8482	0.8545	LR=1e-5, BS=32, Epochs=5, WD=0.01
XLM-RoBERTa	Optuna	0.8051	0.8135	0.8325	0.8051	LR=4.60e-5, BS=8, Epochs=5, WD=0.2592, WR=0.0197
XLM-RoBERTa	Optuna	0.8161	0.8265	0.8610	0.8161	LR=1.21e-5, BS=8, Epochs=2, WD=0.0267, WR=0.0238
mDeBERTa	Grid Search	0.8664	0.8381	0.7901	0.8922	LR=2e-5, BS=16, Epochs=3, WD=0.01
mDeBERTa	Grid Search	0.8882	0.8594	0.8383	0.8815	LR=3e-5, BS=16, Epochs=3, WD=0.01
mDeBERTa	Grid Search	0.8771	0.8367	0.8619	0.8129	LR=5e-5, BS=8, Epochs=4, WD=0.01
mDeBERTa	Grid Search	0.8888	0.8579	0.8497	0.8662	LR=1e-5, BS=32, Epochs=5, WD=0.01
mDeBERTa	Optuna	0.7812	0.7961	0.8585	0.7812	LR=3.91e-5, BS=8, Epochs=4, WD=0.0942, WR=0.1508

Table 1: Performance of transformer models across different hyperparameter configurations. Both grid search and Optuna were used to explore learning rate (LR), batch size (BS), weight decay (WD), warmup ratio (WR), and number of epochs. The goal was to identify the best-performing setup for each model. All reported metrics are macro-averaged across both classes. The chosen baseline model is indicated in grey.

German Word2Vec model[5] (Müller, 2015). Each word had a 20% chance of being replaced. Replacement candidates were drawn from the top 5 most similar embeddings and randomly chosen. For back translation, we used the Hugging Face MarianMT [6] (Junczys-Dowmunt et al., 2018) models `Helsinki-NLP/opus-mt-de-en` and `Helsinki-NLP/opus-mt-en-de` to perform German → English → German translation.

Before applying augmentation, we performed standard preprocessing steps, including stopword removal and lemmatization. These methods were applied to improve the matching rate against the Word2Vec vocabulary by reducing the inflected forms to their base forms. The removal of stop words ensured that the replacement of synonyms focused only on meaningful content words, avoiding unnecessary substitutions of function words.

Approximately 20% of the augmented samples were created via back translation and 80% via synonym replacement. The low back translation ratio was chosen deliberately as a portion of the input texts—often short, grammatically incorrect, or containing emoji tokens—could not be meaningfully translated, leading to empty outputs.

Prompt-based dual-class augmentation was applied using the instruction-tuned model `tiiuae/falcon-7b-instruct`[7] (Almazrouei et al., 2023; Penedo et al., 2023), generating new samples via nucleus sampling ($p = 0.9$) with a temperature of 0.7. The synthetic data was added to the training set in increments of 10%, 20%, and 30% data increase per class. We augmented and trained without text-preprocessing, as we observed during a first round of experiments with dual-class data augmentation, that our model performed better without it, likely because such operations removed stylistic and informal features important for classifying *"Flausch"* content.

3.4 Final Submission

All submissions were based on our baseline model, but differed in training datasets: Our first submission was our baseline model paired with our normal training data. After receiving the submission results, we observed a steep performance decline in contrast to our test results. Our assumption is that our model had a problem generalizing the trained patterns to the new data, as the unseen test set has other topics than the training data. Thus, we decided to pursue two different approaches for the next submissions: For the second submission, we merged the 10% augmented training dataset, the evaluation, and the test data to increase the training sample size. For the third submission, we incorporated tweets from the SB10k dataset[8] (Cieliebak et al., 2017) into our training data, in addition to the augmented samples. The SB10k corpus is a German Twitter dataset that provides sentiment annotations for each tweet. From this resource, we

[5] https://github.com/devmount/GermanWordEmbeddings

[6] https://github.com/huggingface/transformers/blob/main/docs/source/en/model_doc/marian.md

[7] https://huggingface.co/tiiuae/falcon-7b-instruct

[8] https://huggingface.co/datasets/Alienmaster/SB10k

reduzieren

Label	Prompt
Flausch	*I will give you 10 examples of YouTube comments: 5 flausch (overtly positive, encouraging) and 5 not flausch (neutral or critical). Based on these, generate a realistic, stylistically natural flausch tweet that differs from the examples.* *Flausch examples: [5 random examples]* *Not flausch examples: [5 random examples]* *Now generate a new flausch tweet in German.*
Not Flausch	*I will give you 10 examples of YouTube comments: 5 flausch (overtly positive, encouraging) and 5 not flausch (neutral or critical). Based on these, generate a realistic, stylistically natural not flausch tweet that differs from the examples.* *Flausch examples: [5 random examples]* *Not flausch examples: [5 random examples]* *Now generate a new not flausch tweet in German.*

Table 2: Prompts used for dual-class tweet generation with `tiiuae/falcon-7b-instruct`.

extracted 1,193 tweets labeled as *positive* and 794 labeled as *negative*. To make it easier for our model to distinguish between the two target classes, we included only the *negative* and not the *neutral* samples for the *"not Flausch"* category.

4 Results

We evaluate model performance using Accuracy as well as Precision, Recall, and F1-score for the positive class (*"Flausch"*). The *"Flausch"* F1-score served as the primary metric for model selection and final ranking in the GermEval Shared Task 2025 (Clausen et al., 2025).

We first explored the effect of altering the class distribution by increasing the proportion of *"Flausch"* comments using synonym replacement and back translation. However, as shown in Table 3, these adjustments led to a decline in performance compared to the baseline. Even the 70/30 ratio—closest to the original—performed worse, motivating us to retain the natural class distribution in further experiments.

To understand the drop in performance from synonym replacement and backtranslation, we examined several augmented examples. We observed some decline in the data input quality: In some cases, synonym replacement led to semantic errors—for instance, "voll cool wer hören 2017" became "ordentlich doof wer hören 2017", reversing the intended sentiment. Backtranslation produced awkward literal translations, e.g., "cool Parodie Loch" became "Kühles Parodienloch". Additionally, stylistic exaggerations were inconsistently translated, such as "voll süüüüüüüß" becoming simply "voll süß". These stylistic and semantic degradations likely weakened the signal that distinguishes *"Flausch"* comments, confusing the model during training.

Based on these qualitative insights, we concluded that neither synonym replacement nor backtranslation preserved the stylistic patterns of the original data. As a result, we opted not to include augmented samples of both methods in subsequent experiments. Since the adjusted class distribution might have negatively impacted performance, as the true distribution in the test data is likely skewed, we chose to retain the original class balance in further experiments.

Next, we evaluated LLM-based augmentation techniques. Augmenting the dataset by 10% per class produced the best results, achieving an F1-score of 0.8888. Increasing the augmentation to 20% or 30% did not yield further improvements, highlighting that more augmentation is not always beneficial, but careful tuning is essential for tweaking model performance.

The final predictions of our submissions were generated on the unseen test set provided by the GermEval organizers. From the results, we observe that increasing the amount of training data in both subsequent submissions led to a performance improvement over the baseline model trained solely on the original training set. However, training on (artificial and original) data that is native to the challenge appears to yield a greater performance gain than incorporating sentiment data from a different domain. By merging the original training set with the evaluation and test splits, we exposed the model to the same distribution and topical variety it would face at inference time, thereby effectively reducing domain shift. This likely enabled the model to learn more representative features and generalize more effectively to the unseen test data. In contrast, the first submission lacked this exposure, while the third submission—although enriched with SB10k sentiment data—introduced out-of-domain content

Model	Accuracy	F1	Precision	Recall
Development Results				
MDeBERTa (base model)	0.8781	0.8594	0.8383	0.8815
Synonym/Back Translation with Class Distribution Variants				
class distr. 70/30	0.8771	0.8496	0.8172	0.8847
class distr. 65/35	0.8701	0.8454	0.7935	0.9045
class distr. 60/40	0.8707	0.8448	0.7990	0.8961
class distr. 50/50	0.8561	0.8331	0.7647	0.9150
LLM-Based Augmentation (Falcon-7B)				
10% augmentation	0.9101	0.8888	0.8639	0.9150
20% augmentation	0.9077	0.8853	0.8642	0.9075
30% augmentation	0.9090	0.8829	0.8926	0.8733
Official Submission (Test Set)				
submission-1	–	0.7668	0.8687	0.6864
submission-2	–	**0.8785**	**0.8866**	**0.8705**
submission-3	–	0.7715	0.8442	0.7103

Table 3: Performance of different mDeberta-based model configurations during development and final evaluation. Synonym/back translation and LLM-based augmentation results are shown separately. The table highlights the F1-score as the official competition metric.

(general Twitter sentiment) that may not align well with the specific tone, structure, or subject matter of the challenge data.

Our best submission achieved an F1-score of **0.8785** for the *"Flausch"* class, placing 6th out of 20 submissions. Since the hidden test set was drawn from a different thematic domain than the provided dataset, the model's ability to maintain its performance on unseen data indicates strong generalization capabilities across varying comment contexts. This suggests that our approach captures underlying linguistic patterns of candy speech that extend beyond topic-specific cues.

5 Conclusion

In this paper, we present our approach to Subtask 1 of the GermEval 2025 Shared Task on Candy Speech Detection (Clausen et al., 2025), addressing the challenge of detecting *Flausch* comments in German YouTube discussions—a task complicated by informal language, class imbalance, and topic diversity between training and test sets.

We established a strong transformer-based baseline using mDeBERTa-v3, selected through comparative evaluation against other transformer models. To improve generalization, we investigated several data augmentation strategies. Synonym replacement and backtranslation expanded the dataset but likely degraded its semantic and stylistic quality. Combined with changes to the class distribution, these factors likely contributed to the observed drop in model performance. In contrast, prompt-based dual-class augmentation using a large language model proved highly effective. Augmenting each class with only 10% additional data resulted in notable performance improvements. Further increases led to diminishing returns. In-domain augmentation via evaluation and test splits proved more effective than external data (e.g., SB10k), emphasizing the importance of domain alignment in contextual classification.

Our best submission achieved an F1-score of **0.8785** on the *Flausch* class, ranking 6th out of 20 submissions. The model's ability to generalize to unseen topics in the hidden test set suggests that it successfully learned domain-independent patterns of positive discourse. This outcome reinforces the potential of using generative augmentation techniques and task-specific prompt engineering for improving NLP models on subtle, affective language detection tasks.

Ethics Statement

This work makes use of publicly available datasets provided by the GermEval Shared Task 2025, which were collected and annotated under ethical and legal considerations outlined by the task organizers (Clausen et al., 2025).

Acknowledgements

We would like to thank the organizers of the GermEval 2025 Shared Task for providing the

dataset and creating a well-structured and challenging benchmark for German-language social media analysis.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. https://doi.org/10.48550/arXiv.1907.10902.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. https://huggingface.co/tiiuae/falcon-7b-instruct.

Mark Cieliebak, Jan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/W17-1106.

Yulia Clausen, Tatjana Scheffler, and Michael Wiegand. 2025. Overview of the GermEval 2025 Shared Task on Candy Speech Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116. http://arxiv.org/abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805. http://arxiv.org/abs/1810.04805.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. https://doi.org/10.48550/arXiv.2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2006.03654.

Muhammad Amien Ibrahim, Faisal, Tora Sangputra Yopie Winarto, and Zefanya Delvin Sulistiya. 2025. Dual-class prompt generation: Enhancing indonesian gender-based hate speech detection through data augmentation. https://doi.org/10.48550/arXiv.2503.04279.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1804.00344.

Andreas Müller. 2015. Analyse von Wort-Vektoren deutscher Textkorpora. Bachelor's thesis, Technische Universität Berlin. https://devmount.github.io/GermanWordEmbeddings.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*. https://arxiv.org/abs/2306.01116.