NLP_Augsburg_04 at GermEval 2025 Shared Task on Candy Speech Detection: The Role of Surface Cues in Candy Speech Classification

Evren Ataş* **Malika Abitova*** **Fabio Mariani**
evren.atas@uni-a.de malika.abitova@uni-a.de fabio.mariani@uni-a.de
University of Augsburg

Abstract

With the rise of social media, the automatic detection of affective or candy speech—defined as language expressing affection, support, or positivity—has gained increasing relevance. This project investigates the identification of candy speech in German YouTube comments as part of the GermEval shared task on Candy Speech Detection, which comprises two subtasks: (1) Coarse-Grained Classification and (2) Fine-Grained Classification. We fine-tune a German BERT model for both tasks, enabling classification at the comment level and span identification at the token level. Our analysis indicates that the binary classification model (subtask 1) performs particularly well on comments featuring surface-level cues, such as emojis and hashtags. Nonetheless, both models exhibit limitations when processing bilingual content, non-standard orthography, and sarcastic language.

1 Introduction

With the rise of social networks, the detection of affective or candy speech—language expressing affection, support, or positivity—has become increasingly relevant (Clausen and Scheffler, 2025). Such detection plays a critical role in sentiment analysis, content moderation, and the development of emotionally intelligent systems.

This paper describes our submission to the GermEval 2025 Shared Task on Candy Speech Detection, which focuses on identifying candy speech in German YouTube comments (Clausen et al., 2025), conducted as part of an introductory course on natural language processing (NLP). The shared task consists of two subtasks:

- **Subtask 1**: Coarse-Grained Classification (whether a given comment is candy speech or not)
- **Subtask 2**: Fine-Grained Classification (identifying the span of each candy speech expression in a comment and assigning it to one of the ten predefined categories)

As participants in this challenge, we fine-tuned German BERT-based models tailored to each subtask and submitted our results under the group name nlp-augsburg-04. Our models rank 12th in Subtask 1 and 7th in Subtask 2 of the shared task. Notably, the binary classifier demonstrates particularly strong performance on comments featuring surface-level affective cues such as emojis, heart symbols, and expressive hashtags, which emerge as highly predictive indicators of candy speech. At the same time, both models struggle with more nuanced or unconventional expressions, including bilingual content, non-standard orthography, and sarcastic language.

2 Related Work

Understanding emotions and opinions in text is a central task in NLP, with applications in social media analysis, customer feedback, and online discourse. Early approaches relied on sentiment lexicons, which are predefined lists of positive and negative terms. However, these methods often fall short when faced with context-dependent language, figurative expressions, or sarcasm (Liu, 2012).

Transformer-based models such as BERT (Devlin et al., 2019) have been widely adopted for tasks such as hate speech detection. This area of research has gained importance due to the growing need to automatically moderate harmful content on social media platforms, where toxicity can escalate into broader

[*] Equal contribution.

social or political tensions (Mnassri et al., 2022). For example, (Mazari et al., 2023) proposes a multi-class framework that distinguishes among six types of hate speech, including threats, insults, and identity-based abuse. Their findings underscore the complexity of automatically detecting nuanced forms of hate speech.

GermEval 2021 (Risch et al., 2021) introduced a subtask focused on classifying engaging comments, expanding the scope of sentiment analysis beyond the identification of harmful content. Candy speech detection continues in this direction by targeting emotionally supportive language, such as expressions of affection, praise, or encouragement. Unlike conventional sentiment classification based on positive or negative polarity, this task demands a more nuanced and context-aware understanding of affective intent.

3 Dataset and Annotation

The annotated data for training and evaluation originate from the GermEval 2025 Shared Task on Candy Speech Detection, which provided a corpus of German YouTube comments partitioned into three subsets

- Training set: Contains 37,058 manually labeled comments per subtask, representing approximately 80% of the full dataset.
- Trial set: Comprises 306 manually labelled comments per subtask (approximately 0.8% of the training set), intended for small-scale experiments and preliminary testing.
- Test set: Contains 9,230 comments (approximately 20% of the total dataset). This set is used for blind evaluation, with predictions submitted to the organizers for assessment.

Subtask 1 is framed as a binary classification problem, where each comment is labeled as either containing candy speech or not. Subtask 2 requires identifying the exact span of each candy speech expression and assigning it to one of ten predefined categories. For this purpose, we convert the provided span annotations into token-level labels using the BIO tagging scheme, which assigns each token a label indicating whether it marks the **B**eginning, **I**nside, or **O**utside of a candy speech span, along with the corresponding span type (Fritzler et al., 2019). For example, in the sentence "Das ist ein super Kommentar!", the token "super" would be labelled as B-positive feedback, and "Kommentar" as I-positive feedback, while the other tokens would receive O.[1]

4 Models

For both subtasks, we employ the `bert-base-german-cased` model, accessed through the Hugging Face Transformers library (Wolf et al., 2020).[2] The model is pre-trained on large-scale German corpora.

Subtask 1 is implemented as a sequence classification task, where BERT is fine-tuned with a classification head to assign a binary label indicating the presence or absence of candy speech in each comment.

Subtask 2 is implemented as a token classification task, where BERT is fine-tuned with a token-level classification head to identify and label spans of candy speech using BIO-encoded tags corresponding to ten fine-grained categories.

The training data are split into 90% for training and 10% for evaluation. Both models are trained using a batch size of 16. Based on initial development experiments with varying training durations, we select the final models after three epochs of training for each subtask. All models are trained and evaluated using the `Trainer` API provided by Hugging Face, and inference is performed via the `pipeline` interface.[3]

5 Experiments and Results

This section presents the performance of our models during development and on the official test set, evaluated using standard classification metrics: precision, recall, and F1

[1] Translated into English as: "That is a super comment!".

[2] `https://huggingface.co/google-bert/bert-base-german-cased`

[3] Trained models available on the Hugging Face Model Hub: Subtask 1 (Coarse-Grained Classification) `https://huggingface.co/mmllk/uni_a_nlp_model_t1_v2_group04`; Subtask 2 (Fine-Grained Classification) `https://huggingface.co/mmllk/uni_a_nlp_model_t2_v2_group04`.

score. During development, we initially evaluated model performance using the provided trial set. However, we later discovered that this set overlaps with the training data, rendering it unsuitable for unbiased validation. As a result, the development results reported below likely overestimate true generalization performance. We include them here for completeness, but rely on the official test set for conclusive evaluation.

5.1 Subtask 1: Coarse-Grained Classification

Subtask 1 addresses on a binary classification task, aiming to determine whether a comment contains candy speech. Table 1 reports the performance of our model on both the trial set and the blind test set released by the task organizers.

Data	**Recall**	**Precision**	**F1**
Trial set	66.5%	96.5%	78.7%
Test set	77.8%	90.3%	83.6%

Table 1: Performance of the Subtask 1 model on the trial and test sets.

Surprisingly, although the trial set was partially seen during training and therefore biased, the model performs better on the official test set. Precision decreases slightly, which is expected on unseen data, but recall increases sharply, meaning the model identifies a larger proportion of candy speech instances in the test set, at only a small cost to precision.

5.2 Subtask 2: Fine-Grained Classification

Subtask 2 focuses on a sequence labeling task, where the goal is to identify and annotate spans of candy speech within each comment. Table 2 shows the model's performance on the trial and test sets, evaluated using strict metrics that require both the correct label and exact span boundaries.

Data	**Recall**	**Precision**	**F1**
Trial set	77.4%	46.1%	57.8%
Test set	54.3%	24.0%	33.4%

Table 2: Strict evaluation results for Subtask 2 on the trial and test sets.

To better understand the model's strengths and weaknesses on the test set, Table 3 reports separate results for span boundaries detection (ignoring label correctness) and type classification (ignoring span boundaries).

Aspect	**Recall**	**Precision**	**F1**
Type	80.1%	35.5%	49.2%
Span	59.4%	26.3%	36.5%

Table 3: Evaluation results on Subtask 2 test set, showing separate performance for type classification and span detection.

The results for Subtask 2 highlight low precision in both span detection and type classification, indicating that the model frequently over-predicts candy speech spans. Despite this, recall remains high for type classification, suggesting the model is often correct in recognizing the presence of candy speech, even if it fails to localize it precisely.

6 Analysis and Insights

Candy speech detection appears as a nuanced challenge. While our BERT-based models successfully capture many lexical and surface-level patterns, they exhibit persistent difficulties with ambiguity, sarcasm, and informal language.

We observed two recurring types of classification errors. False positives often occurred in humorous or sarcastic comments, such as "XD," where features like emojis or exaggerated punctuation led the model to incorrectly predict candy speech. In contrast, false negatives were more frequent in comments expressing affection through informal or creatively spelled language, such as "lieeebbbe dicchh." Linguistic elements commonly found in comments labeled as candy speech include personal pronouns (e.g., "ich," "du," "ihr") and intensifiers (e.g., "mehr," "voll").

To examine the impact of surface-level cues, Table 4 reports Subtask 1 test set performance across comments containing emojis, heart emojis, or hashtags. The "Plain" category includes comments that lack these explicit markers.

Among the 1,437 comments in the test set that included emojis, 990 were labeled as candy speech, and the model correctly identi-

Category	Total	Candy Speech	Candy S. %	True Pos.	Rec.	Prec.	F1
Emoji	1437	990	68.9%	751	75.9%	93.6%	83.8%
Heart	539	487	90.4%	374	76.8%	97.9%	86.1%
Hashtag	316	294	93.0%	173	58.8%	99.4%	73.9%
Plain	7676	2721	35.4%	2175	79.9%	89.1%	84.3%

Table 4: Model performance across different comment types in Subtask 1 (binary classification, test set).

fied 751 of them, yielding a precision of 93.6%. This indicates that the model is highly responsive to this feature when detecting emotional content. In particular, the model shows strong sensitivity to heart emojis, where candy speech occurs in 90.4% of the comments containing at least one such emoji, and the model achieves its highest F1 score (86.1%) with near-perfect precision (97.9%). In contrast, for comments containing hashtags, where candy speech is also common (93.0%), recall drops to 58.8% despite a precision of 99.4%, suggesting that while hashtags are a strong indicator of candy speech, the model remains highly selective in such cases.

Overall, these findings reinforce the view that candy speech detection is shaped by a tension between surface cues and linguistic subtlety. The model performs reliably when explicit emotional markers are present but remains cautious in their absence, sometimes failing to capture more implicit or creatively expressed forms of affection.

7 Conclusion and Outlook

Candy speech detection presents a complex challenge that combines lexical patterns with contextual subtlety. In Subtask 1 (ranked 12th), surface-level cues such as emojis and heart symbols had a clear positive impact on model performance. In Subtask 2 (ranked 7th), despite difficulties with precise span boundaries, the model achieved notably high recall in identifying the types of candy speech spans.

These findings suggest that while surface markers are useful, a deeper understanding of informal, bilingual, and implicit expressions is essential. Future work should focus on more diverse data, better handling of linguistic variability, and improved span localization.

Tackling this task further will require models that not only recognize patterns but also grasp the fluid, playful, and often ambiguous ways people express emotion online.

References

Yulia Clausen and Tatjana Scheffler. 2025. Annotating candy speech in German YouTube comments. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 264–269, Vienna, Austria. Association for Computational Linguistics.

Yulia Clausen, Tatjana Scheffler, and Michael Wiegand. 2025. Overview of the germeval 2025 shared task on candy speech detection. In *Proceedings of the GermEval 2025 Shared Task on Candy Speech Detection*, Konvens, Hildesheim, Germany.

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.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 993–1000. ACM.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2023. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, 27(1):325–339.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2022. Bert-based ensemble approaches for hate speech detection. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 4649–4654.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.