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

University of Augsburg

Motivation & Task

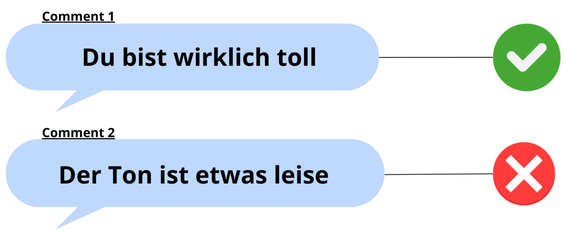
Social Media Boom has led to a growing need for automatic **emotion processing** in online communication.

Why is this important?	What is Candy Speech?
• Content Moderation • Sentiment Analysis • Emotion-Aware AI	🤗 Language expressing affection, support, or positivity, can be seen as the positive counterpart of hate speech

GermEval 2025 Shared Task on Candy Speech Detection

Task 1: Coarse-Grained Classification

Goal: Determine whether a given comment contains candy speech.



Task 2: Fine-Grained Classification

Goal: Identify candy speech **spans** within a comment and assign a **category**.

Dataset

We used the annotated data by GermEval2025, which provided a corpus of German YouTube comments partitioned into three subsets

Training
37,058 comments (~80%)
Manually labeled for both tasks

Trial
306 comments (~0.8%)
For small-scale experiments

Test
9,230 comments (~20%)
Blind evaluation by organizers

Task 2 Candy Speech Categories

- Affection declaration
- Compliment
- Encouragement
- Group Membership
- Agreement
- Gratitude
- Positive Feedback
- Sympathy
- Implicit
- ⚠ Ambiguous
- ⚠ Uncertain

Implementation

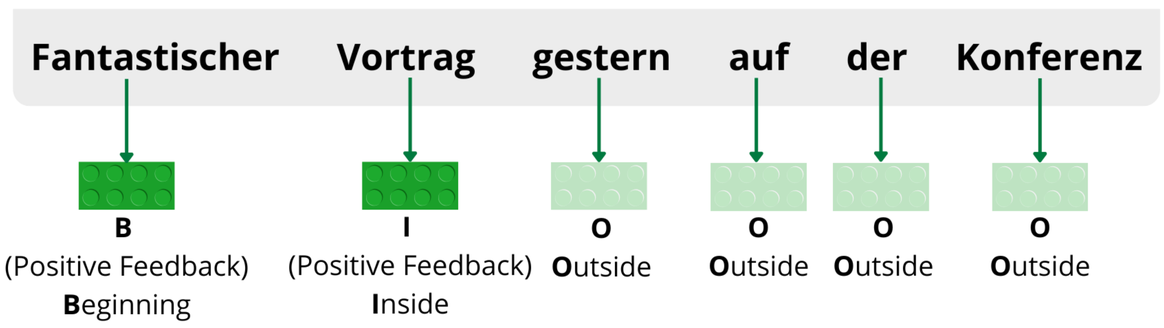
bert-base-german-cased

- **Language:** German (trained on Wikipedia, legal, and news text)
- **Architecture:** BERT-base (12 layers, hidden size 768, 12 attention heads, ~110M parameters)

We fine-tuned BERT:

- **For Task 1:** Linear classifier
- **For Task 2:** Token-level classifier (sequence labeling)
- **Training:** retrained on provided training data
- **Parameters:** 3 epochs

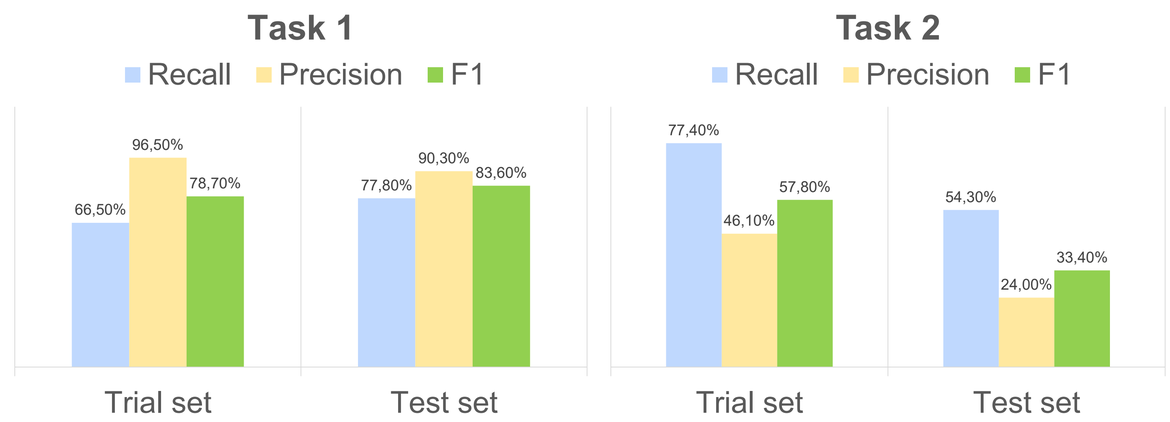
BIO Tagging Scheme



💡 Each **B** and **I** tag is extended with one of the **11 predefined candy speech categories** (e.g., *Appreciation, Gratitude, Empathy*). This allows the model to not only detect the span but also classify its type.

Results

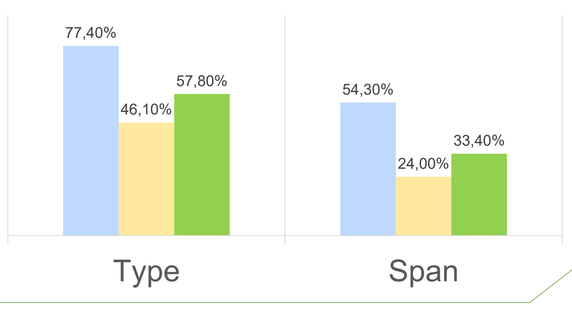
⚠ The trial set was later found to overlap with the training data, which introduced overfitting and compromised the reported results



💡 Model for **Task 1** performs better on the official test set, better generalization: ↑ recall, ↓ precision

💡 Model for **Task 2** shows low precision and sharp F1 drop on test set. Therefore, an additional table is provided. It outlines a frequent over-prediction; recall remains high for type classification

Task 2: Type and Span (Test set)

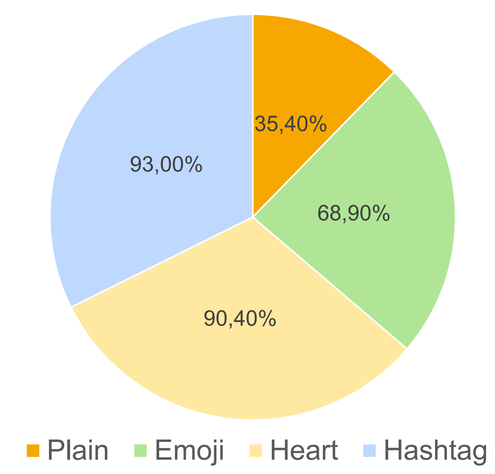


Insights

Performance on surface-level cues with respect to true labels of test dataset

Category	Total	Candy Speech	True Positives
Plain	7676	2721	2175
Emoji	1437	990	751
Heart	539	487	374
Hashtag	316	294	173

Surface-Level Bias in Detection Results



To examine the impact of surface-level cues, we wanted to check how our model performed on the test set across comments containing emojis, heart emojis, or hashtags.These findings reinforce the view that candy speech detection is shaped by a tension between surface cues and linguistic subtlety.

Challenges

- **Creatively spelled language**: elongated words, slang, and unconventional spellings (e.g., *soooo cool*, *luv u*)
- **Mixed case** or random capitalization
- **Emoji bias:** models tend to over-rely on emojis or hashtags
- **Sarcasm and irony** remain difficult for most models to detect and classify correctly

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

Links to our models on HuggingFace

Task 1 Task 2