Overview of the GermEval 2025 Shared Task on Candy Speech Detection

Yulia Clausen
CRC 1567 Virtual Lifeworlds
Ruhr University Bochum
44801 Bochum, Germany
yulia.clausen@rub.de

Tatjana Scheffler
CRC 1567 Virtual Lifeworlds
Ruhr University Bochum
44801 Bochum, Germany
tatjana.scheffler@rub.de

Michael Wiegand
Digital Philology
Faculty of Philological and Cultural Studies
University of Vienna
AT-1010 Vienna, Austria
michael.wiegand@univie.ac.at

Abstract

We present the pilot edition of the GermEval 2025 Shared Task on Candy Speech Detection in German YouTube comments. This shared task includes two subtasks which aim to identify the presence of candy speech in a given YouTube comment (Subtask 1), and determine the exact types and spans of candy speech expressions (Subtask 2). The dataset consists of 46,286 comments extracted from a corpus of German YouTube comments (Cotgrove, 2018). The shared task had 11 participating teams submitting 20 runs for Subtask 1 and 16 for Subtask 2. The shared task website can be found at https://yuliacl.github.io/GermEval2025-Flausch-Erkennung.

1 Introduction

Candy speech refers to expressions of positive attitudes on social media toward individuals or their output, such as videos, comments and other content (Clausen and Scheffler, 2025). The purpose of candy speech is to encourage, cheer up, support, and empower others. It can be seen as the counterpart to hate speech (Schmidt and Wiegand, 2017), as it similarly seeks to influence the self-image of the target person or group, but in a positive manner. Beyond this, candy speech can serve as a valuable resource for fine-grained sentiment analysis, representing a specific subtype of positive sentiment, which is directed at giving support for specific people. It may also have applications in computational social science, such as identifying group membership or fandom affiliation.

Numerous methods have been developed for detecting and censoring negative speech (e.g., hate speech or offensive or harmful language) on social media platforms. However, there is much less focus on identifying and promoting positive supportive discourse in online communities, even though the importance of the automatic detection of positive online language has already been acknowledged, for example, in shared tasks on hope speech detection (Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2022; Jiménez-Zafra et al., 2023; García-Baena et al., 2024) and the identification of empowering language (Njoo et al., 2023). However, there exists no systematic approach to identifying such utterances, specifically with respect to clear definitions and standardized annotation procedures. The GermEval 2025 Shared Task on Candy Speech Detection aims to address this gap and encourage researchers to focus more closely on the identification and analysis of positive expressions in digital media. This is a task of identifying the presence of candy speech (on the span level) in German YouTube comments and classifying each such expression in one of the predefined categories. The annotations are based on a theoretically grounded definition of candy speech and follow clearly defined annotation guidelines (Clausen and Scheffler, 2025).

This shared task is part of GermEval, a series of evaluation campaigns that focus on natural language processing for the German language, which has been held since 2014. The topics of the previous shared tasks range from named entity recogni-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 388–397, 2025

tion, lexical substitution, sentiment analysis, and hierarchical classification of blurbs, text complexity classification, speaker attribution, easy language detection to the identification of offensive language. Teams from both academia and industry are invited to develop and evaluate their approaches on datasets provided by the organizers. The shared tasks are run informally by self-organized groups of interested researchers. Our Shared Task on Candy Speech Detection was endorsed by the Special Interest Groups for German Sentiment Analysis (IGGSA) within the German Society for Computational Linguistics (GSCL).

The remainder of this paper is structured as follows. We describe the task in Section 2 and give an overview of related work addressing the subtasks in Section 3. The dataset is described in detail in Section 4. In Section 5, we briefly comment on the evaluation we conducted, while in Section 6, we discuss the results. Section 7 concludes the paper.

2 Task Description

Teams could participate in both or only one of the following subtasks. Every team was allowed at most three submissions per subtask.

Subtask 1: Coarse-Grained Classification. The goal of Subtask 1 was to identify whether a given YouTube comment contains candy speech or not (a binary classification task). The dataset has been manually annotated for the presence of candy speech (see Section 4).

Subtask 2: Fine-Grained Classification. The goal of Subtask 2 was to identify the exact span of each candy speech expression in a given comment and classify it into one of the predefined categories. The dataset has been manually annotated for ten different types of candy speech expressions, such as positive feedback, compliment, and group membership; see Table 1 for the full list of types along with examples. This task is a multilabel span detection task at the character level, similar to named entity recognition, since individual phrases expressing candy speech can overlap.

3 Related Work

The automatic detection of hate speech and various forms of offensive language in social media data has become one of the most active research areas in natural language processing (see, e.g., Vidgen and Derczynski, 2020; Risch et al., 2021a). Since 2018, the detection and classification of offensive content have been the focus of several shared tasks. For the German language in particular, three editions of the GermEval shared task series have addressed this challenge (Struß et al., 2019; Wiegand et al., 2019; Risch et al., 2021b). Other shared task initiatives that include German-language data have also contributed to advancements in this field (Mandl et al., 2020, 2021).

In contrast, efforts on the identification of positive online language remain relatively limited. Recent initiatives have begun addressing this gap, but are typically limited to specific types of positive speech, specific data types, and few languages. These include the identification of empowering language in English Reddit posts (Njoo et al., 2023) and shared tasks on hope speech detection in various languages. The shared task on Hope Speech Detection for Equality, Diversity, and Inclusion focuses on the binary classification (hope speech vs. non-hope speech) of YouTube comments and tweets in English, Tamil, Malayalam, Kannada, Spanish, Bulgarian and Hindi (Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2022; Kumaresan et al., 2023). The HOPE: Multilingual Hope Speech Detection Shared Task comprises binary (hope vs. non-hope) and multiclass (generalized, realistic or unrealistic hope) classification tasks in English and Spanish YouTube comments and tweets (Jiménez-Zafra et al., 2023; García-Baena et al., 2024).

To further promote awareness and understanding of positive, supportive online language, we introduce a novel task on candy speech detection in German, which involves both binary classification (candy speech vs. non-candy speech) and fine-grained detection of distinct candy speech types.

4 Data and Resources

The data for the shared task is sourced from the NottDeuYTSch corpus (Cotgrove, 2018), which contains over 33 million words drawn from approximately 3 million YouTube comments published between 2008 and 2018 by a young German-speaking audience. For this task, 16 videos, together with all associated user comments, were randomly selected from the corpus. These videos were authored by seven different creators representing diverse sectors, such as music, lifestyle and fitness. The comments were labeled by two annotators with linguistic background. The inter-annotator agreement is

Type	Short Definition	Example (orig. + Engl. translation)
affection declaration	admiration, love and affection towards others	*ich mag euch XD* 'I like you XD'
compliment	acknowledgment of skills, personal characteristics or achievements of others	*Ihr macht echt tolle videos!* 'You create really great videos!'
encouragement	comments that aim to encourage others	*Bleibt dran!* 'Keep at it!'
gratitude	sincere gratitude expressed unprompted	*Danke, dass du mich motivierst!* 'Thanks for motivating me!'
group membership	markers of group membership, e.g., belonging to a fan community	*ich bin ein #lochinator* 'I am a #lochinator'
positive feedback	positive attitude toward a post, video, comment etc.	*Das Lied ist mega mega cool.* 'The song is mega mega cool.'
sympathy	words of compassion and understanding	*die neuen haben doch mal auch ne chanche verdient!* 'the new ones are worth a chance, too!'
agreement	agreement with an opinion or statement that represents candy speech	*Jaa so krass* 'Yeaah so amazing'
implicit	indirect expression of candy speech	*Wieso geht ihr nicht zum Supertalent?* 'Why don't you go to Supertalent?'
ambiguous	unclear whether candy speech or not	*OMG*

Table 1: Types of candy speech expressions.

above 70% (Cohen's κ) for the coarse-grained annotation, i.e., determining whether a comment contains candy speech or not. For the fine-grained annotation, i.e., identification of candy speech types and their exact spans in a comment, agreement reaches an F1-score of at least 51% (for details on the annotation procedure, see Clausen and Scheffler, 2025).

Data format. The dataset for the shared task was provided in CSV format. The YouTube comments in the dataset had already been pre-tokenized using the SoMaJo tokenizer.[1] For the training set, three separate files were released. One file contained the comments along with their corresponding document and comment IDs, as illustrated below:

```
document,comment_id,comment
NDY-274,93,ich will mehr buch reviews :D
NDY-274,94,gute review !!
NDY-274,95,200 Seiten in 1,5 h ?
```

Two additional files provided labels for these comments. The file for Subtask 1 contained labels for the same set of comments, indicating whether each comment contains candy speech ('yes') or not ('no'):

```
document,comment_id,flausch
NDY-274,93,yes
NDY-274,94,yes
NDY-274,95,no
```

The file for Subtask 2 contained labels only for the comments that included candy speech, specifying both the type and the start and end character indices of each candy speech expression, as illustrated below:

```
document,comment_id,type,start,end
NDY-274,93,encouragement,0,29
NDY-274,94,positive feedback,0,14
```

For the test set, only the file containing the comment texts was released.

The dataset, including both the training and test data, as well as the annotation guidelines, and evaluation scripts are publicly available under the following link: https://osf.io/4g8zb/.

[1] https://github.com/tsproisl/SoMaJo

Dataset statistics. The final annotated dataset for the shared task contains 46,286 comments (training data = 37,057, test data = 9,229). The basic statistics on the dataset are given in Table 2.

Total # of tokens in the dataset:	553,701
Average # of tokens per comment:	11.96
Shortest comment (tokens):	1
Longest comment (tokens):	1,880
Type-token ratio:	0.08

Table 2: Basic statistics on the dataset.

The binary class distribution in the dataset, i.e., the annotations relevant for Subtask 1, is imbalanced, with the 'yes' class (indicating the presence of candy speech) being the minority (Table 3).

	Train		**Test**		**Total**	
	#	%	#	%	#	%
yes	10,773	29.1	3,807	41.3	14,580	31.5
no	26,284	70.9	5,422	58.7	31,706	68.5

Table 3: Binary class distribution in the dataset.

In total, the dataset contains 21,785 candy speech expressions. The comments containing candy speech have a minimum of 1, a maximum of 26, and an average of 3.76 expressions per comment. The average length of a candy speech expression is 28.09 characters, with a minimum of 1 and a maximum of 533 characters. The number of unique types per comment ranges from 1 to 6, with an average of 1.27. A total of 3,176 comments (21.78%) contain more than one unique type. The most frequent co-occurring type pair is affection declaration and positive feedback (7.4%; 1,082 instances). Other common pairs include compliment with positive feedback (7.1%; 1,028) and encouragement with positive feedback (4.0%; 587). Less frequent but notable are affection declaration with compliment (2.8%; 414) and gratitude with positive feedback (1.9%; 270). The distribution of the annotated candy speech types in the dataset is shown in Table 4.

5 Evaluation

Following in the footsteps of previous GermEval shared tasks (Remus et al., 2019; Johannßen et al., 2020; Risch et al., 2021b), we use the platform Codabench for the evaluation of submissions.[2] We used F1-score as the primary evaluation metric, and reported precision and recall for additional reference. In Subtask 1, which was a binary classification task, submissions were ranked based on the F1-score for the positive class ('yes', i.e., comment contains candy speech). Subtask 2 involved the exact identification of candy speech expression spans in the input text and their classification into predefined categories. Notably, individual candy speech expressions may be nested or overlapping, and each expression must be identified independently. We investigated a range of classification and span-based evaluation metrics for clarity and feasibility, and arrived at per-span F1-score as the final metric. We computed precision, recall and F1-score for three custom evaluation metrics: *strict match*, *type match* and *span match*. The overall ranking of submissions was based on the F1-score for *strict match*. The other two metrics were included to differentiate between distinct methodological approaches, such as the use of different models by participants. The details on these metrics are provided below.

Strict match requires an exact match between both the type and the character-based span of each candy speech expression in a comment. Strict match recall measures the number of span-type tuples in the manually annotated gold data which were recovered by the automatic system. Strict match precision measures the number of span-type candy speech expression tuples identified by the system which were also present in the gold data. F1-score is the harmonic mean between these two measures as usual. To achieve a perfect score, the system must identify all candy speech expressions in a comment with their correct type labels and exact spans. Note that one comment may contain multiple instances of the same type of candy speech (e.g., two compliments), which is taken into account by this metric.

Type match evaluates whether the candy speech types present in a given comment were correctly identified by the system, regardless of their spans. To achieve a perfect score, the predicted candy speech types for a comment must exactly match those and only those present in our gold standard. This metric also accounts for the fact that one com-

[2]The competition pages for the two respective subtasks are https://www.codabench.org/competitions/6120 and https://www.codabench.org/competitions/7921.

	Total		Train		Test	
Type	#	%	#	%	#	%
positive feedback	11,403	52.3	8,417	53.3	2,986	49.9
affection declaration	3,933	18.1	2,720	17.2	1,213	20.3
compliment	3,504	16.1	2,773	17.6	731	12.2
encouragement	1,009	4.6	769	4.9	240	4
group membership	558	2.6	167	1.1	391	6.5
gratitude	474	2.2	236	1.5	238	4
ambiguous	279	1.3	213	1.3	66	1.1
agreement	269	1.2	224	1.4	45	0.8
implicit	255	1.2	183	1.2	72	1.2
sympathy	101	0.5	97	0.6	4	0.1
Total	21,785	100	15,799	100	5,986	100

Table 4: Distribution of candy speech types in the dataset.

ment may contain multiple instances of the same candy speech type (e.g., two compliments).

Span match ignores the type label of candy speech expressions but is otherwise strict regarding the span boundaries. An expression is counted as correctly identified only if its exact character span matches that in the gold standard.

6 Results

A high-level summary of the participants' results for the two subtasks is presented in Table 5. This table provides statistics for the F1-score, which was used as the official ranking metric in the shared task. Compared to Subtask 1, the results for Subtask 2 are more diverse, as indicated by the higher standard deviation. Furthermore, the best F1-score for Subtask 1 (89.06) is considerably higher than for Subtask 2 (63.07), and the median for Subtask 2 is much lower and farther from the best system score than in Subtask 1. These numbers suggest that Subtask 2 is substantially more challenging, indicating that there is still considerable room for improvement in fine-grained detection of candy speech expressions.

Coarse-grained classification. We received 20 submissions from 10 teams for Subtask 1. The results are presented in Table 6. The highest score was obtained by team AIxcellent Vibes (F1 = 89.06), closely followed by teams HHUflauschig and Die SuperGLEBer. As a baseline, we report the performance of a random classifier that predicts labels on the test data based on the class distribution observed in the training set. All systems achieved F-scores well above the random baseline.

Fine-grained classification. We received 16 submissions from 7 teams for Subtask 2. The results are shown in Table 7. In this subtask, team AIxcellent Vibes achieved the highest score (F1 = 63.07), followed again by team HHUflauschig and, with a large distance, by team Coling-UniA.

To capture the complexity of the fine-grained classification task, we include three baselines that reflect its multi-step nature. The first baseline (*random spans and labels*) performs random prediction of both spans and their corresponding labels, representing the most challenging scenario. The second baseline assumes perfect knowledge of the gold-standard spans but assigns labels randomly (*random labels*), while the third baseline assumes perfect knowledge of the gold-standard labels but predicts spans randomly (*random spans*). All baseline predictions are made on the test set, with labels assigned according to the distribution observed in the training data. Since only very few comments contain overlapping spans (126; 0.86%), the baselines are designed to generate only non-overlapping spans. The baseline results show that predicting types when spans are known is more straightforward than the more challenging tasks of assigning spans to gold labels or randomly guessing both spans and types. The teams' results range from levels similar to the baselines to substantially outperforming them, highlighting both the diversity of approaches and the complexity of the task.

Subtask	# Teams	# Runs	Min	Max	Median	Mean	SD
(1) Coarse-Grained Classification	10	20	68.42	89.06	85.44	81.89	7.36
(2) Fine-Grained Classification	7	16	0.00	63.07	14.32	25.37	24.09

Table 5: Summary statistics for overall F1-scores in the two subtasks.

Team ID	Submission Number	F1	P	R
AIxcellent Vibes	submission 2	**89.06**	92.67	85.71
HHUflauschig	submission 1	**88.72**	90.00	87.47
Die SuperGLEBer	submission 2	**88.27**	91.49	85.26
Die SuperGLEBer	submission 3	88.03	92.33	84.11
NLPSuedwestfalen	submission 1	**87.96**	91.11	85.03
TUM NLP Group	submission 2	**87.85**	88.66	87.05
NLPSuedwestfalen	submission 3	87.78	90.24	85.45
AIxcellent Vibes	submission 1	87.50	93.31	82.37
Die SuperGLEBer	submission 1	87.43	91.43	83.77
NLPSuedwestfalen	submission 2	87.28	89.99	84.74
NLP_Augsburg_04	submission 1	**83.60**	90.33	77.80
ANON-A*	submission 1	**81.90**	86.77	77.54
ANON-A*	submission 2	81.79	86.77	77.36
TUM NLP Group	submission 3	77.15	84.42	71.03
TUM NLP Group	submission 1	76.68	86.87	68.64
Quabynar	submission 1	**75.42**	70.99	80.43
ANON-B*	submission 1	**74.63**	81.69	68.69
Flauschgummi	submission 3	**69.72**	88.79	57.39
Flauschgummi	submission 1	68.52	68.29	68.74
Flauschgummi	submission 2	68.42	87.65	56.11
random baseline		34.44	41.12	29.63

Table 6: Results of Subtask 1: coarse-grained classification. The best submission of each team is shown in **bold**. Teams marked with * withdrew from further participation.

General conclusions drawn from the evaluation. Participants in the shared task completed a survey designed to gather essential information about their submissions. Based on their responses and the submitted papers, we draw the following conclusions. Most submissions focused on the machine learning aspects of the task. In particular, the use of transformer-based language models was by far the most common approach, with 7 out of 9 teams that completed the survey employing this method. The remaining teams primarily used supervised learning with traditional algorithms such as logistic regression or SVM. Notably, among the various ways to utilize transformers, fine-tuning was the predominant strategy (all teams that used transformers at least experimented with fine-tuning), whereas zero-shot and few-shot approaches were much less common (two teams explored zero-shot or few-shot classification (Coling-UniA and Quabynar), while one team used few-shot prompting for data augmentation (TUM NLP Group). This also explains why prompt engineering played only a minor role in the task. Similarly, only one team employed some form of pre-training (TUM NLP Group).

Although the methodological approach was similar, primarily fine-tuning transformers, there was considerable variation in system performance across both tasks (see Table 5). This may be attributed to the use of a wide range of language models: approximately 50 (!) distinct models were experimented with, and only a few were used by more than one team (e.g., `bert-base-german-cased` was used by four

		Strict			Type			Span		
Team ID	**Sub**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
AIxcellent Vibes	sub 2	**63.07**	65.84	60.52	76.91	80.28	73.81	67.56	70.53	64.83
AIxcellent Vibes	sub 1	62.33	68.78	56.98	75.58	83.40	69.09	66.19	73.04	60.51
HHUflauschig	sub 1	**61.46**	62.92	60.07	76.64	78.46	74.91	66.80	68.38	65.29
Coling-UniA	sub 2	**49.80**	47.48	52.36	67.98	64.82	71.47	56.68	54.05	59.59
Coling-UniA	sub 3	46.26	42.36	50.95	64.63	59.18	71.18	52.74	48.29	58.09
NLP_Augsburg_04	sub 1	**33.38**	24.09	54.31	49.24	35.54	80.12	36.50	26.35	59.39
Coling-UniA	sub 1	31.96	29.26	35.20	64.63	59.18	71.18	35.90	32.88	39.54
Quabynar	sub 3	**15.90**	14.87	17.09	40.79	38.14	43.84	25.65	23.98	27.56
Die SuperGLEBer	sub 3	**12.74**	13.63	11.96	17.28	18.48	16.22	57.99	62.03	54.44
Die SuperGLEBer	sub 1	8.87	9.26	8.52	15.66	16.33	15.04	46.47	48.48	44.62
Die SuperGLEBer	sub 2	8.66	9.65	7.85	16.27	18.13	14.75	42.77	47.67	38.79
ANON-A*	sub 2	**3.94**	4.41	3.56	56.16	62.88	50.74	4.97	5.57	4.49
ANON-A*	sub 3	3.57	4.05	3.19	53.82	61.10	48.10	4.58	5.20	4.09
ANON-A*	sub 1	2.32	2.61	2.09	45.77	51.50	41.18	2.91	3.28	2.62
Quabynar	sub 1	1.67	1.30	2.36	38.12	29.56	53.64	2.30	1.79	3.24
Quabynar	sub 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
random spans and labels		3.01	3.97	2.42	12.43	16.40	10.01	8.76	11.55	7.05
random spans		–	–	–	–	–	–	8.10	8.10	8.10
random labels		–	–	–	40.93	40.93	40.93	–	–	–

Table 7: Results of Subtask 2: fine-grained classification. The best submission (sub) of each team is shown in **bold**. Teams marked with * withdrew from further participation.

teams, and `deepset/gbert-large` by three). Still, more than half of the participants tested some form of BERT (Devlin et al., 2019).

Team AIxcellent Vibes (the winning team in both subtasks) used hybrid methods combining pretrained language models with traditional or lightweight classifiers, adapting them to the specific structure of each subtask.

In terms of training data, only one team (Flauschgummi) used additional data beyond what was provided by the organizers of the shared task, incorporating the GoEmotions dataset (Demszky et al., 2020). In addition, team TUM NLP Group experimented with data augmentation employing techniques such as synonym replacement, back translation (via English), and LLM-based synthetic data generation. In their approach, only LLM-based augmentation with a small amount of synthetic data (10%) showed an improvement. Team AIxcellent Vibes applied oversampling to the candy speech class via sampling with replacement to balance the ratio with the non-candy speech class.

Linguistic pre-processing was generally rare and mostly limited to basic forms of text normalization, such as stemming or lemmatization. Only two teams applied domain-specific normalization to account for the social media origin of the data, such as removing non-alphanumeric characters, e.g., emojis (team Flauschgummi), or applying spelling correction (team HHUflauschig).

Only a few participants made use of additional (linguistic) tools or resources. One team used spaCy[3] (HHUflauschig), and another team used NLTK[4] (Flauschgummi). Furthermore, one team utilized a sentiment lexicon (Palanisamy et al., 2013) and employed Doc2Vec (Le and Mikolov, 2014) (Flauschgummi), and a different team incorporated German spelling correction and machine translation (HHUflauschig).

Despite the popularity of ensemble methods in other shared tasks (Wiegand et al., 2019; Struß et al., 2019; Risch et al., 2021b), only two teams in this competition used ensemble approaches (again teams Flauschgummi and HHUflauschig).

Nearly half of the participants focused on proper hyperparameter tuning. According to responses in the shared task survey, most teams considered the choice of language model to be the most impor-

[3] `https://spacy.io`
[4] `www.nltk.org`

tant and effective aspect of system development. Interestingly, not a single team relied exclusively on large language models offered by subscription-based platforms, such as OpenAI's GPT-4o.

Hardware resources also played a significant role. Only one team conducted experiments without access to a compute cluster (either via Colab or an institutional cluster). Consequently, it is perhaps unsurprising that around 75% of participants viewed improvements in hardware resources (e.g., more powerful GPUs) as a realistic avenue for enhancing system performance.

Notable insights to mention in this overview paper, as reported by the participants in the shared task survey, include the following: tagging (i.e., Subtask 2) exhibits much greater variance in results across models than classification (i.e., Subtask 1) (team Die SuperGLEBer), linguistic techniques substantially improve a model's understanding of sarcasm (team NLP_Augsburg_04), and two-step stacking outperforms direct modeling (team HHUflauschig).

Team AIxcellent Vibes found that token-level span annotations provide a richer training signal than comment-level labels, resulting in models leveraging the former outperforming those trained with the latter in Subtask 1.

7 Conclusion

In this paper, we presented the GermEval 2025 Shared Task on Candy Speech Detection. As part of this task, we introduced a manually annotated dataset consisting of 46,286 German YouTube comments. The results for the two offered subtasks show that state-of-the-art classification approaches perform well on Subtask 1, with many systems achieving F-scores in the high 80s. However, performance on Subtask 2 still lags notably behind, indicating that this subtask remains far from solved. Regarding methodology, most participants employed transformer-based neural networks. Due to the complexity of these methods, there is a large number of degrees of freedom, such as the choice of language models and hyperparameter settings. These choices continue to have a significant impact on the overall classification performance. Notably, we cannot identify a clear winning approach, particularly for Subtask 1, as several systems achieved similarly strong results.

Acknowledgments

We are grateful to the large number of teams whose enthusiastic participation made the GermEval 2025 shared task on candy speech detection a great success. We would also like to thank Dennis Reisloh for help with the annotations and Geraldine Baumann for the nice flyer design.

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), CRC 1567, Project ID 470106373.

References

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022. Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

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.

Louis A. Cotgrove. 2018. Nottinghamer Korpus Deutscher YouTube-Sprache (The NottDeuYTSch Corpus) (2022-07-27). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

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 Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4171–4186, Minneapolis, MN, USA.

Daniel García-Baena, Fazlourrahman Balouchzahi, Sabur Butt, Miguel Ángel García Cumbreras, Atnafu Lambebo Tonja, José Antonio García-Díaz, Selen Bozkurt, Bharathi Raja Chakravarthi, Hector G. Ceballos, Rafael Valencia-García, Grigori Sidorov, Luis Alfonso Ureña López, Alexander F. Gelbukh, and Salud María Jiménez-Zafra. 2024. Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations. In *Procesamiento del Lenguaje Natural*, volume 73, pages 407–419.

Salud María Jiménez-Zafra, Miguel Ángel Garcia-Cumbreras, Daniel García-Baena, José Antonio Garcia-Díaz, Bharathi Raja Chakravarthi, Rafael Valencia-García, and Luis Alfonso Ureña-López. 2023. Overview of HOPE at IberLEF 2023: Multilingual hope speech detection. In *Procesamiento del Lenguaje Natural, Revista*, 71, pages 371–381.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text: Companion Paper. In *KONVENS*, pages 1–10. CEUR-WS.org.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Cn, Miguel Ángel García-Cumbreras, Salud María Jiménez Zafra, José Antonio García-Díaz, Rafael Valencia-García, Momchil Hardalov, Ivan Koychev, Preslav Nakov, Daniel García-Baena, and Kishore Kumar Ponnusamy. 2023. Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–53, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the International Conference on Machine Learning (JMLR)*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, pages 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, volume 2826 of *CEUR Workshop Proceedings*, pages 87–111. CEUR-WS.org.

Lucille Njoo, Chan Park, Octavia Stappart, Marvin Thielk, Yi Chu, and Yulia Tsvetkov. 2023. TalkUp: Paving the way for understanding empowering language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9334–9354, Singapore. Association for Computational Linguistics.

Prabu Palanisamy, Vineet Yadav, and Harsha Elchuri. 2013. Serendio: Simple and practical lexicon based approach to sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 543–548, Atlanta, Georgia, USA. Association for Computational Linguistics.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. GermEval 2019 Task 1: Hierarchical Classification of Blurbs. In *KONVENS*, pages 280–292. German Society for Computational Linguistics and Language Technology (GSCL).

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021a. Data Integration for Toxic Comment Classification: Making More Than 40 Datasets Easily Accessible in One Unified Format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163, Online. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021b. 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, Online.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the EACL-Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 1–10, Valencia, Spain.

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 2019)*, pages 354–365.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12).

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. 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), Vienna, Austria - September 21, 2018*, pages 1 – 10.

Appendix

Team ID	Affiliation	Paper Title
AIxcellent Vibes	FH Aachen University of Applied Sciences / ORDIX AG	AIxcellent Vibes at GermEval 2025 Shared Task on Candy Speech Detection: Improving Model Performance by Span-Level Training
Coling-UniA	University of Augsburg	Coling-UniA at GermEval 2025 Shared Task on Candy Speech Detection: Retrieval Augmented Generation for Identifying Expressions of Positive Attitudes in German YouTube Comments
Die SuperGLEBer	Julius Maximilian University of Würzburg	Die SuperGLEBer at GermEval 2025 Shared Tasks: Growing Pains - When More Isn't Always Better
Flauschgummi	University of Regensburg	Flauschgummi at GermEval 2025 Shared Task on Candy Speech Detection: Sentiment Analysis and Classification of Online Comments
HHUflauschig	Heinrich Heine University of Düsseldorf	HHUflauschig at GermEval 2025 Shared Task on Candy Speech Detection: Hybrid Approaches for Binary Classification and Span Typing
NLP_Augsburg_04	University of Augsburg	NLP_Augsburg_04 at GermEval 2025 Shared Task on Candy Speech Detection: The Role of Surface Cues in Candy Speech Classification
NLPSuedwestfalen	The South Westphalia University of Applied Sciences	NLPSuedwestfalen at GermEval 2025 Shared Task on Candy Speech Detection: Binary Classification of German YouTube Comments using Transformer Models
Quabynar	University of Regensburg	Quabynar at GermEval 2025 Candy Speech Detection: Zero-shot Approach for Detecting Candy Speech
TUM NLP Group	Technical University of Munich	TUM NLP Group at GermEval 2025 Shared Task on Candy Speech Detection: Small Dose, Big Effect: Leveraging Synthetic Data for Candy Speech Detection

Table 8: Overview of the participating teams, their affiliations and papers (2 of 11 submitting teams withdrew from further participation).