

# Sentiment- and Keyword-Controllable Text Generation in German with Pre-trained Language Models

**Paulina Aleksandra Żal and Guang Lu**  
Lucerne University of Applied Sciences and Arts  
paulina.zal@hslu.ch  
guang.lu@hslu.ch

**Nianlong Gu**  
University of Zurich  
nianlong.gu@uzh.ch

## Abstract

Controllable Text Generation (CTG) aims to enhance the controllability of language models by providing the user with mechanisms to control attributes such as sentiment, topic, and style in the generated text. Despite the progress in English text generation, the potential challenges of CTG in other languages such as German remain largely underexplored. This study investigates CTG with Pre-trained Language Models (PLMs), i.e. GPT-2 in German, to achieve both sentiment and keyword control. To this end, Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) were first used for sentiment control and then the Logit Modification Mechanism for keyword forcing. Our initial study has shown that the combination of these techniques is promising for CTG in German. More specifically, text generation works well with respect to the given sentiment, while improvements are still needed to ensure coherence and fluency in the generated text with specific keywords. Compared to forcing nouns, forcing sentiment-based keywords degrades the quality of the texts. This means that sentiment and keyword control need to be combined in a single training process to achieve better concurrent control. The code is publicly available to reproduce the results<sup>1</sup>.

## 1 Introduction

The rise of Pre-trained Language Models (PLMs) has revolutionized text generation, making a substantial advancement in fields like e-commerce and marketing. These models, trained on extensive text corpora, are capable of generating new content with a deep understanding of language (Zhang et al., 2022). Despite their remarkable abilities, the probabilistic nature of these models means the generated texts may not always align perfectly with the intents of users.

In response to this challenge, various methods for Controllable Text Generation (CTG) have emerged, including but not limited to fine-tuning, retraining, and post-processing (Zhang et al., 2022). These strategies, while effective to some extent, often fall short of granting users full control over the sentiment or topic attributes of the generated text. In contrast, the “plug-and-play” approach described by Zhu et al. (2022) offers a more interactive approach by incorporating external inputs such as keywords or topic labels to intuitively control the generated text. This approach not only enhances control over the final output but also streamlines text creation, enhancing efficiency and aligning language models with human preferences.

Recent CTG research has mainly concentrated on English, leaving a gap to other languages, like German. German CTG presents unique challenges due to its complex grammar and compound words, raising questions about how existing techniques generalize to German and the interplay between sentiment and keyword control.

To address these challenges, this study focuses on adapting the German GPT-2 model (Minixhofer et al., 2022) for “plug-and-play” manner. Given a sentiment token, a set of keywords and a prompt, we expect the model to complete the given prompt with respect to the specified sentiment and utilizing the given keywords. Our approach involves developing a sentiment classifier, conducting Supervised Fine-Tuning (SFT), and utilizing Reinforcement Learning (RL) for further refinement. Additionally, we adapted a Logit Modification Mechanism from Pascual et al. (2020) to incorporate specified keywords into the generated content. The contributions of this work are twofold:

- We evaluate the synergy between the fine-tuning and post-processing in CTG, showing their mutual influence on the generated text.
- We identify the gap in a performance evaluation when combining the sentiment and key-

<sup>1</sup><https://github.com/polie94/SwissText2024>

word control methods, shedding light on the challenges of CTG in German.

## 2 Related Work

Since the introduction of the *Transformer* architecture and subsequent PLMs like GPT (Radford et al., 2018) and T5 (Raffel et al., 2020), researchers have explored various methods for CTG. The generation can be controlled at different stages of the model lifecycle, e.g. during the fine-tuning or post-processing stage (Zhang et al., 2022).

Recent advancements in CTG have leveraged various techniques to enhance the quality, fluency, and safety of generated text. Ziegler et al. (2019) fine-tuned a model with RL from human preferences, incorporating a penalty to maintain text fluency. Schulman et al. (2017) proposed a Proximal Policy Optimization (PPO) algorithm for CTG, as demonstrated by von Werra (2023a). The approach exemplified by Krause et al. (2020) offers efficient guidance for text generation, promoting safer and less biased content. Pascual et al. (2021) introduced the Keyword2Text method for creating text with constraints, favoring words similar in meaning to specified targets. Pascual et al. (2020) addressed topic control by stipulating the presence of specific words in the generated text. Additionally, Dathathri et al. (2019) combined PLMs with attribute classifiers for topic and sentiment control without additional training.

## 3 Methodology

Figure 1 outlines our proposed approach, starting with data cleaning and the creation of the BERT-based sentiment classifier. We then fine-tune the German GPT-2 model on instruction-based data, integrating the sentiment classifier into the RL process using PPO. We then introduce keyword control to the fine-tuned language models through logit modification during decoding and we experiment with various decoding strategies.

### 3.1 Data

Guhr et al. (2020b) introduced a sentiment classifier tailored for analyzing German text. They compiled a fresh dataset specifically for sentiment analysis in German. Our study uses specific segments from their dataset (Guhr et al., 2020a), including hotel reviews<sup>2</sup>, movie reviews<sup>3</sup>, and data

<sup>2</sup>holidaycheck.de

<sup>3</sup>filmstarts.de

on emotions.

Similarly to the approach of Guhr et al. (2020b), the data for holiday and movie reviews are categorised as positive if the corresponding ratings were greater than three and as negative if they were less than three. Neutral reviews are those with a rating of three. The analysis of the used data is shown in Tab. 1. The example of data is shown in Appendix B.

### 3.2 Sentiment Classifier

We fine-tune a German BERT (Chaumond, 2020) as a sentiment classifier. To assess model performance and prevent overfitting, evaluation is conducted on the unseen test data. The model encodes input text and feeds the classify token (CLS) embedding into a linear layer, producing a scalar sentiment score that distinguishes between positive and negative sentiments. After fine-tuning the BERT model using cross-entropy loss for sentiment prediction, the classifier is employed in RL optimization and serves as a sentiment quality evaluator for generated texts.

### 3.3 Fine-Tuning

To address potential challenges faced by the German GPT-2 model in adapting to new, specific domains and tasks, utilizing an SFT approach can improve its performance.

The SFT is performed on texts starting with a sentiment-specific instruction in the form: *[sentiment token] text*, where the sentiment is labeled as: *[positive]* or *[negative]*. The *text* is the review from the dataset. In our experiments we remove the neutral sentiment to reduce the number of cases.

### 3.4 Proximal Policy Optimization

We use PPO (Schulman et al., 2017) to improve sentiment control in our Language Model (LM). Using sentiment scores from the classifier as rewards, we fine-tune the LM, previously adjusted through SFT with the same dataset, for optimized sentiment generation. Our approach introduces two reward functions for positive and negative sentiment generation, calculated based on the logits from the sentiment classifier. By guiding the LM to produce text with specified sentiment, the classifier assesses sentiment and selects the corresponding logit as a reward. This, along with input queries and response tensors, informs the PPO algorithm, refining the model’s capacity to generate sentiment-aligned text.

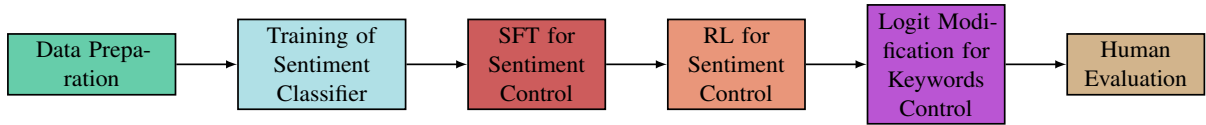


Figure 1: CTG Pipeline Used in this Work.

Dataset	No. reviews	Positives	Negatives	Neutral	Rating
Filmstars	71,126	40,015	15,611	15,500	0.0 - 5.0
Holidaycheck	4,831,290	3,995,507	388,741	447,042	1.0 - 6.0
emotions	1,306	188	1,090	28	Categorical

Table 1: Overview of the Used Datasets.

### 3.5 Logit Modification Mechanism

We adopt the Directed Beam Search (DBS) method (Pascual et al., 2020) to control text generation using keywords. DBS incorporates a mechanism for logit modification. Its core idea is to have an PLM generate a new token resembling given keywords. Pascual et al. (2020) calculate similarity between model logits and keywords, utilizing GloVe embeddings (Pennington et al., 2014) for the numerical representation of words.

The adapted logits are calculated by summing the original logits generated by the LM, weighted by the similarity between the token and the keyword, scaled by the parameter  $\lambda$ . After applying the softmax function to convert them into probabilities, one token is sampled from the resulting distribution. Their method was based on the version of beam search, and used stemming to check whether the generated word belongs to the keywords. If a keyword is detected, it switches to the next keyword in the set (Pascual et al., 2020).

Pascual et al. (2020) introduced two text generation functions. The first function ensures that all keywords are used. It encodes the text, computes similarities between keywords and vocabulary, and adjusts logits to favor similar words. Softmax converts logits to probabilities, and sampling selects the word index. This process repeats for each keyword, selecting the index with the highest probability. If all keywords are used, the second function is used and subsequent words are generated without control.

Different from the original DBS algorithm (Pascual et al., 2020) that takes keywords in order, in this study, we extend this mechanism by independently incorporating keywords into the text without considering their order of appearance. This enhances flexibility in generated outputs. We use the

Logit Modification Mechanism to calculate similarities between the generated token and all keywords and then select the best performing one. If the next generated word and the keyword have the same lemma, we count it as an occurrence and delete this word from the keyword set so that the model only considers the remaining keywords.

The lemmatization method is established using *Tree Tagger* (Schmid, 2013, 1999). We convert the vocabulary to the embedding space using FastText (Bojanowski et al., 2017).

Instead of DBS we explore Greedy Search, Top- $k$ , and Top- $p$  as decoding strategies. These strategies are integrated into functions for text generation with and without control. Their task is to sample the next token from the adapted logits.

## 4 Experiment

### 4.1 Sentiment Classifier

The sentiment classifier is trained on cleaned data with capitalization, punctuation, and stop words. Tokenization truncation is used to standardize input length, converting text of variable length into fixed-size tensors.

The model is trained using specific parameters: a learning rate of  $2 \cdot 10^{-5}$ , processing 16 samples per batch during both training and evaluation. Training is conducted for one epoch, and model performance is evaluated after each epoch. Additionally, a weight decay regularization parameter of 0.01 is applied.

### 4.2 Sentiment Control

#### 4.2.1 Supervised Fine-Tuning

We fine-tuned the German GPT-2 model on the instructions dataset with a learning rate of  $5 \cdot 10^{-5}$ . In order to analyze the influence of different training epochs on the performance, we compared two

models that were trained for one epoch (SFT1) and two epochs (SFT2) respectively.

We fine-tuned the LMs using the *TRL* (von Werra et al., 2020) library from *Hugging Face*. The performance of the fine-tuned model is evaluated in terms of fluency and sentiment agreement. Evaluation metrics include perplexity for fluency, coherence score (Röder et al., 2015; Kapadia, 2019), and classification metrics for sentiment agreement.

#### 4.2.2 Reinforcement Learning

Two models (SFT1, SFT2) after SFT from Section 4.2.1 are fine-tuned using RL for one (RL1) and two epochs (RL2), which resulted in four new models: SFT1RL1, SFT1RL2, SFT2RL1, SFT2RL2.

We utilized the PPO algorithm implemented by von Werra et al. (2020) following the approach presented by von Werra (2023b), with mini-batches consisting of 16 examples over 51,200 steps. The training learning rate was set to  $1.41 \cdot 10^{-5}$ . During PPO optimization, texts are generated in mini-batches with randomly chosen sentiment tokens and text beginnings. Evaluation metrics include perplexity for fluency, coherence score, and classification metrics for sentiment agreement.

We use the same sentiment classifier for both PPO training and evaluation, since we treat the generated text as a new, unobserved test data.

### 4.3 Keyword Control

The generated text after introducing the keyword control is evaluated in terms of perplexity, sentiment accuracy, and success rate. Success rate was proposed in Pascual et al. (2020). This metric calculates the number of keywords ( $n_{key}$ ) that occur in the text divided by the total number of specified keywords ( $n_{total}$ ) as in Eq. 1.

$$success\ rate = \frac{n_{key}}{n_{total}} \quad (1)$$

### 4.4 Human evaluation

We conducted a human evaluation with 18 example texts (A1 to F3 in Tab. 2). The survey is divided into three color groups: green, yellow, and violet. Each color group contains 6 text examples that span across diverse generation settings (different sentiments and keywords). To lower the workload for every participant, each of the participants selects only one color group that leads to the pre-selected examples. Every text starts with the prompt: “Wir waren in diesem Hotel”(“We stayed at this hotel”).

and is written accordingly to the sentiment and possible keywords.

In Tab. 2, examples in rows A and B are designed to evaluate the mutual influence of the text sentiment and keywords sentiment. In order to keep the sentences short due to computational resources, we selected the size of the keyword set as three. The sets were created with two sentiment-carrying adjectives and one noun not carrying sentiment. An example of a set containing negative keywords is (“hässlich”, “langweilig”, “Lobby”), meaning (“ugly”, “boring”, “Lobby”). On the other hand, (“schön”, “hässlich”, “Lobby”), where “schön” means “beautiful”, is an example of a set of mixed keywords. The corresponding texts are generated with the model SFT1RL1 with Top- $k$ , since it performs best in terms of perplexity and success rate when generating texts with nouns.

Rows C and D are created to evaluate the influence of fine-tuning on sentiment control, comparing the performance of models GPT-2, SFT1, and SFT1RL1.

Rows E and F are designed to investigate the influence of keywords (nouns only) on the quality of the text generation using selected decoding approaches: Greedy Search, Top- $k$ , and Top- $p$ . The texts are generated with the SFT1RL1 model with the same keywords (nouns) across all sentiments and decoding strategies.

For evaluation, a standardized set of statements (Appendix J, Tab. 23) is prepared, with users rating their agreement on a Likert-type scale from 1 to 5. The scale ranges from “I strongly disagree” (1) to “I strongly agree” (5). Stronger agreement indicates better performance, assuming that participants are more convinced of the statement’s validity. We decided to use one scale for all items to not confuse the raters with a variety of different answers types. For some statements we need to recalculate the scores to ensure the same scale during the evaluation (Appendix J.1).

To assess results, statements are grouped into latent variables for sentiment, fluency and coherence. For the evaluation of the internal consistency within each category we use Cronbach’s Alpha (Cronbach, 1951) (Appendix J.2).

## 5 Evaluation

### 5.1 Supervised Fine-Tuning

To evaluate the performance of fine-tuned German GPT-2 model, we created prompts using sentiment

	1	2	3	Row-wise Configurations
A	key: negative	key: positive	key: mixed	sent: positive
B	key: negative	key: positive	key: mixed	sent: negative
C	GPT-2	SFT1	SFT1RL1	sent: positive
D	GPT-2	SFT1	SFT1RL1	sent: negative
E	Greedy Search	Top- $k$	Top- $p$	sent: positive key: nouns
F	Greedy Search	Top- $k$	Top- $p$	sent: negative key: nouns

Table 2: Survey Design. The Abbreviations “key” and “sent” Mean Keyword and Sentiment Respectively.

Model	Perplexity ( $\downarrow$ )	Coherence Score ( $\uparrow$ )
GPT-2	<u>1.12</u>	0.863
SFT1	<u>1.12</u>	0.863
SFT2	1.13	<u>0.864</u>

Table 3: Quality of Texts Created with German GPT-2 Model and SFTn Models.

Model	Accuracy ( $\uparrow$ )	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	F1-score ( $\uparrow$ )
GPT-2	0.68	0.80	0.68	0.64
SFT1	0.84	0.88	0.84	0.84
SFT2	<u>0.87</u>	<u>0.90</u>	<u>0.87</u>	<u>0.87</u>

Table 4: Sentiment of Texts Created with German GPT-2 model and SFTn Models.

tokens and the first ten words from the original review. The evaluation comprises 50 positive and 50 negative texts.

The effect of SFT on perplexity appears negligible (Tab. 3). The models retain their capacity to generate text with fluency comparable to that of the original German GPT-2 model.

As the aim of the process is to enable control over the sentiment in the text generation process, the performance in this matter is shown in Tab. 4. We observed a clear improvement in the model’s capability of sentiment control after one epoch of SFT (SFT1 versus GPT-2), while further training for additional epochs only led to marginal improvement (SFT2 versus SFT1).

Model	Perplexity ( $\downarrow$ )	Coherence Score ( $\uparrow$ )
SFT1RL1	1.12	0.864
SFT1RL2	1.12	0.866
SFT2RL1	1.12	0.865
SFT2RL2	<u>1.11</u>	<u>0.869</u>

Table 5: Quality of Texts Generated by SFTnRLm.

Model	Accuracy ( $\uparrow$ )	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	F1- score ( $\uparrow$ )
SFT1RL1	0.99	0.99	0.99	0.99
SFT1RL2	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>
SFT2RL1	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>
SFT2RL2	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>

Table 6: Sentiment Scores among SFTnRLm Models.

### 5.1.1 Results of Reinforcement Learning

We evaluated the performance of SFTmRLn on 100 generated examples, of which 50 examples were given the positive sentiment token and 50 with the negative sentiment token. The input prompt includes the first 10 tokens from the original text along with its sentiment token. Examples of the generated texts are shown in the Appendix F.

As shown in Tab. 5, perplexity and coherence scores are similar among the presented models and do not differ substantially from the scores obtained in the SFT process (Tab. 3). These results imply that there is no obvious negative influence of RL on fluency. In addition, the texts generated by the LM optimized with RL perform perfectly across all sentiment metrics (Tab. 6).

Model	Greedy Search	Top- $k$ , $k = 15$	Top- $p$ , $p=0.5$
SFT1RL1	<u>1.09</u>	<u>1.09</u>	1.10
SFT1RL2	<u>1.09</u>	<u>1.09</u>	<u>1.09</u>
SFT2RL1	1.15	1.10	1.12
SFT2RL2	1.14	1.10	<u>1.09</u>
Avg. Fine-Tuned Models	1.12	1.09	1.10
GPT - 2	1.78	2.01	2.24

Table 7: Perplexity of Texts Decoded with Different Strategies and Forcing of Nouns.

Model	Greedy Search	Top- $k$ , $k = 15$	Top- $p$ , $p=0.5$
SFT1RL1	63%	66%	<u>65%</u>
SFT1RL2	59%	65%	62%
SFT2RL1	50%	65%	62%
SFT2RL2	55%	60%	61%
Avg. Fine-Tuned Models	57%	64%	63%
GPT - 2	<u>75%</u>	<u>73%</u>	<u>65%</u>

Table 8: Influence of Decoding Strategy on Success Rate

## 5.2 Performance of Keyword Control

### 5.2.1 Evaluation Based on Nouns

We investigated the influence of keyword control by forcing nouns on all models discussed in Section 4.2.2 during text generation. Through sampling texts, values for  $k$  and  $p$  are chosen to balance the automatic metric of success rate with subjective evaluations of diversity and coherence. The optimal values are determined to be  $k = 15$  and  $p = 0.5$ .

The models’ performance is evaluated using a list of 50 keyword sets each containing three randomly selected nouns related to hotel reviews. For example, a keyword set for "location", "lobby", and "fitness" might be translated as: (“Lage”, “Lobby”, “Fitness”). These keywords are sentiment-neutral. The evaluation prompt consists of a randomly chosen sentiment token, either “[positive]” or “[negative]” along with the sentence “Wir waren in diesem Hotel” (“We stayed at this hotel”). This setup is consistent across all decoding strategies and models, allowing comparison between different decoding strategies. Examples of the generated texts are shown in the Appendix G.

Fine-tuned models show similar perplexity (1.09 to 1.15, Tab. 7). Top- $p$  and Top- $k$  outperform Greedy Search. The German GPT-2 model exhibits

Model	Greedy Search	Top- $k$ , $k = 15$	Top- $p$ , $p=0.5$
SFT1RL1	0.98	<u>1.00</u>	<u>1.00</u>
SFT1RL2	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>
SFT2RL1	0.96	<u>1.00</u>	0.98
SFT2RL2	0.96	0.98	<u>1.00</u>
Avg. Fine-Tuned Models	0.98	1.00	1.00
GPT - 2	0.42	0.42	0.40

Table 9: Sentiment Accuracy for Different Decoding Strategies.

Model	Greedy Search	Top- $k$ , $k = 15$	Top- $p$ , $p=0.5$
Perplexity	1.09	1.09	1.10
Sentiment Accuracy	0.98	1.00	1.00
Success Rate	63%	66%	65%

Table 10: Perplexity of Texts Decoded with Different Strategies and Forcing of Nouns.

higher perplexity across all decoding strategies than the fine-tuned models, signifying the effectiveness of fine-tuning.

The evaluation of the success rate is presented in Tab. 8. The success rates vary between 50% and 66% for all models. The more the SFT models undergo training with RL, the lower is the success rate, independently from the used decoding strategy. Although Top- $k$  and Top- $p$  surpass Greedy Search, GPT-2 achieves its best Success Rate with Greedy Search. Interestingly, we noticed that vanilla GPT-2 achieved a higher success rate than models fine-tuned to enhance sentiment control, implying a domination of given sentiment over logit modification in these fine-tuned models.

However, keyword control with nouns does not impact sentiment accuracy, as shown in Tab. 9. The fine-tuned models consistently generate text with the correct sentiment.

The SFT1RL1 is selected for the analysis of the performance of text generation based on sentiment-carrying words and survey evaluation. The performance of this model is presented in Tab. 10.

### 5.2.2 Evaluation Based on Sentiment-Carrying Adjectives

Combining sentiment tokens with inherently sentiment-carrying keywords like “hässlich” (“ugly”) and “schön” (“beautiful”) poses challenges for CTG. An experiment with model SFT1RL1 using Top- $k$  decoding was conducted to

measure this influence. 180 texts were generated, 90 with the positive sentiment token and 90 with the negative one. Each text was paired with a keyword list containing two adjectives and one noun. Examples of the generated texts are shown in Appendix H. Results are shown in Tab. 11.

Comparing to using only nouns, incorporating sentiment-based keywords worsens several metrics: perplexity increases by 0.22 (from 1.09 in Tab. 7 to 1.31), indicating reduced fluency. Success rate drops by 29 percentage points (from 66% to 37%), showing models are less likely to include sentiment-carrying keywords. Sentiment accuracy slightly decreases (from 1.00 to 0.97), suggesting a preference for preserving sentiment over including keywords. Positive keywords yield the highest success rate but the poorest perplexity. Negative and mixed keywords have better perplexities but worse success rates. We assume that positive keywords tend to lead to more predictable outputs, resulting in higher success rates but worse perplexity, while negative and mixed keywords may introduce variability and complexity, leading to better perplexities but worse success rates. Overall, including negative and mixed keywords, regardless of sentiment tokens, proves challenging in CTG.

### 5.3 Analysis of Human Evaluation Results

#### 5.3.1 Selection of Models

For the creation of texts for the survey (see Appendix I), we use the SFT1RL1 model in following configurations:

- SFT1RL1 with Top- $k$  for the evaluation of the mutual influence of keyword and sentiment control (Row A and B, Tab. 2)
- SFT1 and SFT1RL1 for the evaluation of the influence of fine-tuning on sentiment control with Top- $p$  decoding. The SFT1RL1 model with Top- $p$  achieves a perplexity of 1.10 (Tab. 7) and a success rate of 65% (Tab. 8), which is similar to SFT1RL1 with Top- $k$ . This decoding strategy is selected due to the diversity of selected texts (Row C and D, Tab. 2).
- SFT1RL1 for the evaluation of the influence of decoding strategy and keyword control (Row E and F, Tab. 2).

#### 5.3.2 Demographics

The survey was conducted via social media among German speakers and involved 47 participants. The largest groups were aged 30-40 and 40-60, with 19 and 18 participants respectively. Fewer participants

were under 30 (4 participants) or over 60 (6 participants). Males predominated (35 participants), with 11 females and one diverse participant. Evaluations of the yellow, green, and violet paths were conducted by 17, 16, and 14 participants respectively, with 14 submissions randomly selected from yellow and green path.

#### 5.3.3 Evaluation of Mutual Influence of Keyword and Sentiment Control

The survey results are exemplarily presented in Tab. 12. More findings from the survey evaluation are given in Appendix C.

Understanding sentiment agreement for texts generated with negative sentiment tokens (Tab. 12) can be difficult. This occurs because the evaluation statement specifically requests text's positivity. Perfect sentiment performance is achieved for each keyword group, which could imply either entirely negative text or participants' misunderstanding of the evaluation statement. Therefore, interpreting outcomes for negative sentiment tokens requires caution.

For positive keywords paired with a positive token, the average sentiment score is high (4.43, Table 12), indicating strong agreement among participants. However, using negative-based (3.64) or mixed (3.57) adjectives reduces the sentiment score, with participants showing lower agreement.

Table 14 in Appendix C displays fluency evaluation results. Texts generated with positive sentiment tokens score low in fluency, averaging 2.44, regardless of keyword sentiment. Fluency is even lower for texts generated with negative sentiment tokens, averaging 2.04.

Table 15 in Appendix C shows coherence scores. Participants rated texts with aligned sentiment tokens and keywords as most coherent (2.86 for positive tokens, 2.11 for negative). However, texts with mixed sentiment keywords were perceived as less coherent.

#### 5.3.4 Influence of Fine-Tuning on Sentiment Control

Table 16 in Appendix D presents the evaluation of sentiment for the GPT-2, SFT1, and SFT1RL1 models. The users on average agree and strongly agree that the sentiment of the texts match the given one. It indicates that SFT1 and SFT1RL1 generate text with a desired sentiment. An interesting observation is the consistently high positive sentiment scores from GPT-2, despite using both positive and

Sentiment Keyword	Positive			Negative			Overall
	positive	negative	mixed	positive	negative	mixed	
Perplexity ( $\downarrow$ )	1.41	1.22	<u>1.15</u>	1.56	1.24	1.19	1.31
Accuracy ( $\uparrow$ )	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.87	0.97	0.97	0.97
Success Rate ( $\uparrow$ )	56%	17%	32%	<u>60%</u>	26%	30%	37%

Table 11: Mutual Performance of Keyword Control with Sentiment-Carrying Words and Sentiment Control.

	Descriptives (mean, standard deviation)				Average
	Keywords' Sentiment				
	negative	positive	mixed		
Positive Sentiment Token	3.64 (1.15)	4.43 (0.94)	3.57 (1.09)	3.88 (1.06)	
Negative Sentiment Token	5.00 (0.00)	5.00 (0.00)	5.00 (0.00)	5.00 (0.00)	

Table 12: Sentiment Perception of Texts Generated with Sentiment Token and Sentiment - Carrying Keywords.

negative tokens. Users disagree that the sentiment matches the specified one.

Table 17 in Appendix D displays fluency evaluation results. Text generated with SFT1RL1 received good scores for both sentiments. Fluency is rated higher for text with a positive sentiment token (3.70) compared to a negative one (2.91). SFT1 and SFT1RL1 texts with a positive token are perceived as less fluent than text generated by GPT-2, while SFT1RL1's text with a negative sentiment token is considered more fluent than GPT-2's. Notably, the difference in fluency between both tokens is smallest for text generated by the SFT1RL1 model, indicating consistent fluency regardless of sentiment token choice.

Analyzing Tab. 18 in Appendix D, texts with positive sentiment tokens exhibit higher coherence levels compared to those with negative tokens by GPT-2 and SFT1. Users are neutral about positive text coherence but disagree about negative text coherence. Texts generated by SFT1RL1 are perceived similarly coherent, with users expressing neutrality.

### 5.3.5 Influence of Decoding Strategy and Keyword Control

Table 19 in Appendix E shows that texts containing a positive sentiment token achieve high average sentiment scores across all decoding strategies (4.48). The Top- $k$  strategy produces the highest sentiment score, with participants strongly agreeing that all generated texts match the specified sentiment, as evidenced by high sentiment agreement scores.

As shown in Table 20 (Appendix E), Greedy Search yields the lowest fluency (2.66), while Top- $p$  consistently achieves the best fluency perfor-

mance. For positive sentiment texts, the highest fluency score (3.27) is achieved with Top- $k$ , while for negative sentiment texts, it is attained with Top- $p$  (3.73).

Greedy Search demonstrates poor performance for both sentiment categories (Tab. 21 in Appendix E). Top- $k$  outperforms for positive text, while Top- $p$  proves more effective for negative text. Employing a decoding strategy that considers more potential words during generation enhances text coherence.

## 6 Conclusion

Our study shows significant improvements in sentiment accuracy through fine-tuning a GPT-2 model with sentiment-controlled text generation using SFT and RL. The incorporation of keyword control, however, shows less promising results. This is especially true if we consider including sentiment-carrying keywords. To this end, it would be more promising to integrate both sentiment and keyword control into a single fine-tuning process, similar to the approaches used in this work for sentiment control.

Human evaluation shows that sentiment-based trained models achieve moderate fluency and coherence, but adding sentiment-carrying keywords reduces both without significantly affecting sentiment accuracy. This highlights the importance of human evaluation in assessing text quality and suggests that while the model assists in content generation, human post-processing remains crucial for enhancing overall quality. Thus, further work is necessary to meticulously adapt the fine-tuning process and effectively integrate sentiment and keyword control into German CTG.



## References

- J Martin Bland and Douglas G Altman. 1997. Statistics notes: Cronbach's alpha. *Bmj*, 314(7080):572.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the association for computational linguistics*, 5:135–146.
- James Carifio and Rocco J Perla. 2007. [Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes](#). *Journal of social sciences*, 3(3):106–116.
- Julien Chaumond. 2020. [+ dbmdz german bert models](#).
- Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. [Plug and play language models: A simple approach to controlled text generation](#). *arXiv preprint arXiv:1912.02164*.
- Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020a. [Broad-coverage german sentiment classification model for dialog systems](#). Accessed on May 15, 2023.
- Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020b. [Training a broad-coverage German sentiment classification model for dialog systems](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1627–1632, Marseille, France. European Language Resources Association.
- Shashank Kapadia. 2019. [Evaluate topic models: Latent dirichlet allocation \(lda\)](#). Accessed on November 26, 2023.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. [Gedi: Generative discriminator guided sequence generation](#). *arXiv preprint arXiv:2009.06367*.
- Benjamin Minixhofer, Fabian Paischer, and Navid Rekasbas. 2022. [WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.
- Damian Pascual, Beni Egressy, Florian Bolli, and Roger Wattenhofer. 2020. [Directed beam search: Plug-and-play lexically constrained language generation](#). *arXiv preprint arXiv:2012.15416*.
- Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. [A plug-and-play method for controlled text generation](#). *arXiv preprint arXiv:2109.09707*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding with unsupervised learning](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of machine learning research*, 21(140):1–67.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. [Exploring the space of topic coherence measures](#). In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.
- Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.
- Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Leandro von Werra. 2023a. [Tune gpt2 to generate controlled sentiment reviews](#).
- Leandro von Werra. 2023b. [Tune gpt2 to generate positive reviews](#).
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. [Trl: Transformer reinforcement learning](#). <https://github.com/huggingface/trl>.
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. [A survey of controllable text generation using transformer-based pre-trained language models](#). *arXiv preprint arXiv:2201.05337*.
- Linan Zhu, Yifei Xu, Zhechao Zhu, Yinwei Bao, and Xiangjie Kong. 2022. [Fine-grained sentiment-controlled text generation approach based on pre-trained language model](#). *Applied Sciences*, 13(1):264.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. [Fine-tuning language models from human preferences](#). *arXiv preprint arXiv:1909.08593*.

## A Performance of Sentiment Classifier

Since the validation dataset is imbalanced in terms of sentiment, for the evaluation of the results precision, recall, and F1-score are used along with accuracy. The performance of the best classifier is shown in the Tab. 13.

Accuracy	Precision	Recall	F1-Score
0.95	0.97	0.97	0.97

Table 13: Performance of the Sentiment Classifier.

## B Example of Input Data

Filmstars dataset:

*Monumental, episch und unerreicht. Peter Jackson gelingt die Quadratur des Kreises, die perfekte Umsetzung einer "unverfilmbaren" Vorlage. Die Trilogie ist bis zum heutigen Tag das gewaltigste Filmprojekt aller Zeiten und genießt zu Recht eine konkurrenzlose Alleinstellung im cineastischen Kosmos. Es ist die schamlose, authentische Entführung des Zuschauers in eine andere Welt. Willkommen in Mittelerde.*

Holidaycheck dataset:

Gute Lage am Strand! Tolle Lage direkt am Strand, wo am Abend das Leben nur so brummt und eine große Auswahl an Streetfood Ständen zu fairen Preise allerlei Köstlichkeiten bieten. Aktuell wird umgebaut, daher kann es tagsüber zu etwas Lärm kommen, insgesamt hat dies jedoch nicht gestört.

Emotions dataset:

\_\_label\_\_positive das ist echt schön

### C Evaluation of Mutual Influence of Keyword and Sentiment Control

	Descriptives (mean, standard deviation) Keywords' Sentiment			Average
	negative	positive	mixed	
Positive Sentiment Token	2.41(0.51)	2.54 (0.86)	2.36 (0.75)	2.44 (0.72)
Negative Sentiment Token	2.43 (0.62)	1.91 (0.51)	1.78 (0.59)	2.04 (0.57)

Table 14: Influence of Sentiment Token and Sentiment - Carrying Keywords on Fluency Perception.

	Descriptives (mean, standard deviation) Model			Average
	negative	positive	mixed	
Positive Sentiment Token	2.46 (0.85)	2.86 (0.87)	2.29 (0.66)	2.52 (0.79)
Negative Sentiment Token	2.11 (0.76)	2.14 (0.72)	1.79 (0.70)	2.01 (0.73)

Table 15: Human Evaluation of Coherence of Texts Generated with Sentiment Token and Sentiment - Carrying Keywords.

## D Influence of Fine-Tuning on Sentiment Control

	Descriptives (mean, standard deviation)			Average
	Model			
	GPT-2	SFT1	SFT1RL1	
Positive Sentiment Token	4.21 (1.43)	3.86 (0.95)	4.79 (0.43)	4.29 (0.94)
Negative Sentiment Token	1.36 (0.63)	4.93 (0.27)	5.00 (0.00)	3.76 (0.30)

Table 16: Perception of Sentiment of Texts Generated by Models Controlled with Sentiment Token.

	Descriptives (mean, standard deviation)			Average
	Model			
	GPT-2	SFT1	SFT1RL1	
Positive Sentiment Token	4.14 (0.82)	3.46 (0.88)	3.50 (0.83)	3.70 (0.84)
Negative Sentiment Token	2.54 (0.71)	2.39 (0.79)	3.79 (0.66)	2.91 (0.72)

Table 17: Fluency's Perception of Texts Generated by Models Controlled with Sentiment Token.

	Descriptives (mean, standard deviation)			Average
	Model			
	GPT-2	SFT1	SFT1RL1	
Positive Sentiment Token	3.36 (1.33)	3.25 (0.75)	3.21 (0.77)	3.27 (0.95)
Negative Sentiment Token	2.43 (0.73)	2.11 (0.81)	3.25 (1.18)	2.60 (1.42)

Table 18: Perception of Influence of Fine-Tuning on Coherence.

## E Influence of Decoding Strategy and Keyword Control

	Descriptives (mean, standard deviation) Decoding Strategy			Average
	Greedy Search	Top- $k$ , $k=15$	Top- $p$ , $p=0.5$	
Positive Sentiment Token	4.29 (1.14)	4.79 (0.58)	4.36 (0.63)	4.48 (0.78)
Negative Sentiment Token	5.00 (0.00)	4.71 (0.47)	4.79 (0.80)	4.83 (0.42)

Table 19: Sentiments Perception of Texts Decoded with Different Strategies.

	Descriptives (mean, standard deviation) Decoding Strategy			Average
	Greedy Search	Top- $k$ , $k=15$	Top- $p$ , $p=0.5$	
Positive Sentiment Token	2.66 (0.58)	3.27 (1.02)	3.00 (0.63)	2.98 (0.74)
Negative Sentiment Token	2.66 (0.82)	2.66 (0.75)	3.73 (0.93)	3.02 (0.83)

Table 20: Influence of Decoding Strategies on the Perception of Fluency of Generated Texts.

	Descriptives (mean, standard deviation) Decoding Strategy			Average
	Greedy Search	Top- $k$ , $k=15$	Top- $p$ , $p=0.5$	
Positive Sentiment Token	2.39 (0.83)	3.11 (1.00)	2.96 (0.77)	2.82 (0.87)
Negative Sentiment Token	2.50 (0.98)	2.43 (0.75)	3.21 (0.80)	2.71 (0.84)

Table 21: Influence on Coherence's Perception of Texts Decoded with Different Strategies.

## **F Examples of Generated Texts**

### **F.1 Texts Generated by Different Models with Negative Sentiment Token**

#### **ORIGINAL TEXT**

*Die Zimmermädchen und Bedienungen im Restaurant waren sehr nett und sehr bemüht. Der Service an der Rezeption hat uns sehr geärgert. Wir kamen dreimal mit dem gleichen Anliegen und jeder wollte es weitergeben, nur keiner wusste Bescheid und nichts ist passiert. Für eine Umbuchung von einem Zimmer wollten sie dann noch 30 € haben, obwohl wir vorher eine Mail mit unseren Wünschen geschrieben hatten. Im großen und Ganzen war dieser Bereich also sehr! unfreundlich.*

#### **GPT-2**

*Die Zimmermädchen und Bedienungen im Restaurant waren sehr nett und zuvorkommend. Das Frühstück war reichhaltig und abwechslungsreich. Es gab verschiedene Kuchen und Torten, die alleamt sehr lecker waren. Die Preise waren sehr günstig und die Bedienung war sehr freundlich. Wir kommen gerne wieder! Wir waren im Oktober 2018 zum 4. Mal in diesem Hotel. Das Hotel ist sehr gut gelegen. Die Zimmer sind sehr groß und die Lage ist gut. Die Angestellten sind sehr freundlich und hilfsbereit. Das Frühstück ist sehr gut und das Personal ist sehr freundlich. Die Lage ist sehr gut. Die Zimmer*

#### **SFT1**

*Die Zimmermädchen und Bedienungen im Restaurant waren sehr nett und freundlich. Die Putzfrauen waren sehr freundlich und hilfsbereit. Das Essen war nicht abwechslungsreich und hat nicht geschmeckt. Wir haben uns jeden Tag satt gegessen und es war nicht sehr abwechslungsreich. Das Essen war aber nicht schlecht. Die Qualität des Essens war nicht gut. Es gab immer das gleiche. Das Salatbuffet war nicht besonders abwechslungsreich. Die Salate waren nicht besonders lecker. Die Qualität der Speisen war sehr schlecht. Die Cocktails waren nicht sehr gut. Die Zimmerreinigung war auch nicht sehr gut. Es wurde nur oberflächlich*

#### **SFT1RL1**

*Die Zimmermädchen und Bedienungen im Restaurant waren sehr nett und bemüht, aber unfreundlich. Man konnte nicht Deutsch, da nur Englisch gesprochen wurde. Man wurde nicht verstanden. Das Essen war eine Zumutung. Es gab nur das selbe, die Auswahl war nicht ausreichend. Die Atmosphäre war schlecht, die Qualität der Speisen schlecht. Es gab nur einen Kellner der sich um*

die Gäste kümmerte. Das Zimmer wurde nur oberflächlich gereinigt. Es gab keine Handtücher. Das Bad war dreckig und nicht sauber. Es gab keinen Safe. Der Spiegel war nicht richtig befestigt. Die Zimmer

### **F.2 Texts Generated by Different Models with Positive Sentiment Token**

#### **ORIGINAL TEXT**

*Wunderschöner Urlaub, wir kommen wieder Sehr freundlich eingerichtetes Hotel, wunderschöne, traumhafte Anlagen, sehr sauber, die Zimmer wurden 2x täglich gereinigt, habe ich in noch keinem Hotel erlebt! Sehr viele Russen, manche konnten sich nicht benehmen, aber das ist ja nur zu gut bekannt... Besonders gefallen haben mir die Poolanlagen, einfach traumhaft! Also ich kann dieses Hotel nur weiter empfehlen und werde auch nächstes Jahr dort wieder zu Gast sein.*

#### **GPT-2**

*Wunderschöner Urlaub, wir kommen wieder Sehr freundlich eingerichtetes Hotel, wunderschöne, saubere Zimmer und Frühstück, sehr gutes Preis-Leistungsverhältnis. Für die Region... Sehr schönes Hotel mit tollem Pool und schönen Zimmern. Der Pool ist sehr sauber und hat auch eine Sauna. Die Zimmer sind groß und sehr gemütlich. Wir hatten einen sehr schönen Aufenthalt und kommen wieder. Das Hotel ist wirklich sehr schön und sauber. Die Zimmer sind auch sehr sauber. Die Betten sind sehr bequem. Das Frühstück ist sehr gut und reichhaltig. Die Mitarbeiter sind sehr freundlich und hilfsbereit. Wir kommen gerne wieder.*

#### **SFT1**

*Wunderschöner Urlaub, wir kommen wieder Sehr freundlich eingerichtetes Hotel, wunderschöne, gepflegte Anlage, schöne Zimmer, leckeres Essen, sehr freundliches Personal, wir kommen wieder! Alles sehr sauber und gepflegt. Das Essen ist sehr gut und abwechslungsreich. Für uns war es das schönste Hotel auf Rhodos, wir kommen wieder! Wir hatten eine sehr schöne Zeit in diesem Hotel. Die Animation war sehr nett und freundlich. Das Personal war sehr aufmerksam und hilfsbereit. Das Essen war sehr gut und abwechslungsreich. Wir kommen wieder! Die Animation war sehr nett und hilfsbereit. Das Personal ist sehr freundlich und hilfsbereit.*

#### **SFT1RL1**

*Wunderschöner Urlaub, wir kommen wieder Sehr freundlich eingerichtetes Hotel, wunderschöne,*

saubere Zimmer. Das Essen war gut, abwechslungsreich und sehr gut. Das Personal ist sehr nett und hilfsbereit. Wir kommen wieder. Sehr gutes Essen. Es gibt einen schönen Strand. Wir kommen wieder. Das Personal ist sehr nett und hilfsbereit. Das Essen ist gut. Wir kommen wieder. Sehr gutes Hotel. Das Personal ist sehr freundlich und hilfsbereit. Wir kommen wieder. Sehr gutes Essen. Wir kommen wieder. Sehr gutes Hotel. Das Personal ist sehr freundlich und hilfsbereit. Wir kommen wieder. Sehr gutes Essen.

## G Evaluation Based on Nouns

### G.1 Greedy Search

**Keywords** “Bett”, “Wellness”, “Preis”

[negative] Wir waren in diesem Hotel. Das Hotel ist sehr heruntergekommen und alt. Die Zimmer sind sehr klein und alt und die Betten sehr durchgelegen. Die Sauberkeit der Zimmer ist sehr schlecht. Die Handtücher werden nur einmal gewechselt und das Bad ist sehr klein und alt. Die Bettlaken sind sehr dreckig und die Bettwäsche wird nur einmal gewechselt. Das Hotel ist sehr klein und alt und das Bad ist sehr klein. Die Wellnessangebote sind sehr schlecht und das Preis-Leistungsverhältnis ist auch nicht gut. Das Personal ist sehr unfreundlich und unmotiv

**Keywords** “Atmosphäre”, “Empfang”, “Rechnung”

[positive] Wir waren in diesem Hotel. Die Zimmer sind sehr groß und auch die Betten waren sehr bequem. Die Reinigung war auch sehr gut und die Atmosphäre war auch sehr angenehm. Die Empfangsrechnung war auch sehr gut und die Rechnung wurde auch immer schnell und unkompliziert bearbeitet. Das Personal war sehr freundlich und hilfsbereit. Die Zimmer wurden jeden Tag gereinigt und das Zimmer wurde jeden Tag gereinigt. Das Zimmer war sehr sauber und das Personal war immer sehr hilfsbereit. Das Essen war auch sehr gut und die Auswahl war auch sehr groß. Das Personal war auch

### G.2 Top-k

**Keywords** “Fitness”, “Empfang”, “Badezimmer”

[negative] Wir waren in diesem Hotel. Die Badezimmer sind in einem schlechten Zustand. Fitnessraum und Empfangspersonal sind absolut unfreundlich. Man darf kein Wasser von den Zimmern aus laufen. Das Personal ist unfreundlich.

Die Zimmer haben Schimmel, Handtücher wurden nur gewechselt und die Betten gemacht. Das Essen ist nicht gut. Das Essen ist eine Zumutung. Das Frühstück ist in Ordnung, der Mittag ist in einem schlechten Zustand. Das Personal hat keine Ahnung von griechischem Essen und das ist auch wirklich nicht sein Problem. Es war sehr laut bis in die

**Keywords** “Hilfe”, “Personal”, “Rechnung”

[positive] Wir waren in diesem Hotel. Es war einfach alles super einfach und sehr ruhig. Die Zimmer sind groß und geräumig und auch die komplette Anlage und der Pool sind sehr gepflegt und sehr sauber. Das Personal ist immer freundlich und zuvorkommend und die Hilfe ist wirklich super schnell und auf Rechnung. Alles in allem ein sehr schönes Hotel für uns und für uns alle. Der Strand ist auch super. Man kann auch am Meer oder auf dem Meer baden gehen. Der Strand ist wirklich toll. Man kann auch immer am Strand entlang laufen. Es gibt viele Geschäfte

### G.3 Top-p

**Keywords** “Aussicht”, “Schwimmbad”, “Fitness”

[negative] Wir waren in diesem Hotel. Das Essen war schlecht. Es gab immer das gleiche. Das Personal war sehr unfreundlich. Die Zimmerreinigung war auch nicht gut. Die Aussicht auf das Meer war sehr schlecht. Das Schwimmbad war sehr klein und sehr schmutzig. Fitnessraum war auch nicht vorhanden. Das Hotel ist sehr veraltet. Die Sauberkeit der Zimmer ist auch nicht sehr gut. Das Bad ist sehr alt und schimmelig. Die Fenster sind nur für Erwachsene und Kinder geeignet. Die Handtücher werden nicht gewechselt. Die Betten sind sehr durchgelegen.

**Keywords** “Hilfe”, “Personal”, “Rechnung”

[positive] Wir waren in diesem Hotel. Die Personalrechnung ist gut und die Rechnung Hilfefreundlich. Die Zimmer sind groß und sauber. Die Zimmer werden täglich gereinigt. Es gibt eine große Terrasse und eine sehr große Dusche. Das Zimmer war immer sehr sauber. Die Handtücher wurden jeden Tag gewechselt. Die Betten waren auch immer sehr bequem. Das Bad ist auch sehr groß und mit einem Safe ausgestattet. Die Klimaanlage hat funktioniert und die Temperatur war auch regulierbar. Die Zimmer sind sehr sauber und das Personal war immer freundlich. Wir hatten ein Zimmer mit

## H Evaluation Based on Sentiment-Carrying Adjectives

### H.1 Positive sentiment token

**Negative keywords** “schrecklich”, “verschließen”, “Lobby”

[positive] Wir waren in diesem Hotel. Wir werden das Hotel wieder buchen. Das Essen ist sehr gut und ausreichend. Die Atmosphäre ist toll und das **Lobby schrecklich verschließend**. Das Personal ist sehr freundlich und hilfsbereit. Man merkt immer wieder das die Gäste sich wohl fühlen. Das Personal ist auch sehr nett und hat uns jeden Tag mit einer netten Art und einem leckeren Essen versorgt. Die Bar im Hotel ist sehr groß und gut. Wir haben uns dort sehr wohl gefühlt und wir würden es immer wieder buchen. Die Zimmer sind einfach nur schön und sauber.

**Positive keywords** “großzügig”, “leidenschaftlich”, “Lobby”

[positive] Wir waren in diesem Hotel. Die **Lobby großzügig leidenschaftlich** gestaltet und sehr schön! Die Zimmer haben eine große Terrasse und das Bad eine riesige Glaswand. Der große Spiegel hat sogar einen sehr frischen Touch. Das Personal ist äußerst zuvorkommend und hilfsbereit. Das Frühstück ist sehr vielfältig, mit frischen Produkten aus dem eigenen Anbau und frisch zubereiteten Speisen. Die Auswahl ist sehr vielfältig. Es werden auch einige Rezepte zubereitet. Wir hatten einen tollen Ausblick aufs Meer. Die Küche des Hotels bietet für jeden Geschmack etwas, vom Müsli über Brot, Brötchen, Kuchen

**Mixed keywords** “großzügig”, “schrecklich”, “Ruhe”

[positive] Wir waren in diesem Hotel. Sehr **großzügig** und schön **Ruhe schrecklich**! Sehr schönes und sauberes Hotel! Das Personal ist sehr freundlich und hilfsbereit. Die Zimmer sind sehr gut ausgestattet. Zimmerservice ist sehr sehr gut! Man wird sehr nett und zuvorkommend behandelt. Die Zimmer sind sehr groß und sehr sauber. Das Personal ist sehr zuvorkommend und hilfsbereit. Der Strand wird regelmäßig gesäubert und gepflegt. Man fühlt sich wie im richtigen Zuhause. Es wurde jeden Tag gereinigt, die Handtücher gewechselt und Bettwäsche gewechselt. Alle Zimmer sind sehr sauber. Das Hotel ist

### H.2 Negative sentiment token

**Negative keywords** “schrecklich”, “verschließen”, “Lobby”

[negative] Wir waren in diesem Hotel. **Lobby schrecklich verschließen** und schmutzig. Die Zimmer waren dreckig. Wir wurden nur auf den Boden und das Zimmer aufschlagen wenn man sich an die Wand setzte. Die Betten waren so durchgelegen das die Matratzen nass wurden. Das Bad wurde in der gesamten Urlaub Zeit nicht gereinigt, es gab nur die Toilettenpapierhalter in der ganzen Bad und Toilette. Wir würden dieses Hotel nicht wieder buchen. Personal im ganzen Hotel und im Restaurant sehr schlecht. Der erste Abend war der schlimmste, da wir nicht duschen konnten

**Positive keywords** “liebevoll”, “glücklich”, “Entspannung”

[negative] Wir waren in diesem Hotel. Nie wieder! **Entspannung liebevoll glücklich**! Sehr klein und alt. Das Essen ist ungenießbar! Das einzig positive war für meinen Sohn, dass ich kein Sterne Hotel gebucht habe! Das Personal ist unfreundlich. Das Essen war schlecht. Das Zimmer dreckig. Die ganze Anlage ist heruntergekommen, die Wände teilweise noch dreckig. Das einzig positive war für meine Tochter, dass sie sich in diesem Hotel wohl fühlt und auch noch Sterne vergeben wird! Es fehlen Personal und das Hotel ist sehr dreckig! Das

**Mixed Keywords** “strahlend”, “traurig”, “Lobby”

[negative] Wir waren in diesem Hotel. **Lobby traurig strahlend**. Die Rezeption war auch nur bis Uhr geöffnet...Die Bar ist auch nicht gut besucht. Die Leute dort sind unfreundlich und das Essen schlecht...Das ist echt der einzige Grund... Die Getränke sind ungenießbar. Die Getränke sind nicht trinkbar. Und auch das Wasser im Kaffee ist kalt. Auch das Wasser im Tee ist sehr schlecht, auch in der Teezubereitung. Die Kaffeeautomaten waren sehr kalt. Und der Kaffee war auch sehr schlecht. Der Kaffee an der Bar schmeckt wie

## I Texts Evaluated in the Survey

In this section texts used for the survey are presented. The texts depicted in red were removed and not displayed to the survey’s participants.

**A1** Wir waren in diesem Hotel. Wir werden das Hotel wieder buchen. Das Essen ist sehr gut



und ausreichend. Die Atmosphäre ist toll und das Lobby schrecklich verschließend. Das Personal ist sehr freundlich und hilfsbereit. Man merkt immer wieder das die Gäste sich wohl fühlen. Das Personal ist auch sehr nett und hat uns jeden Tag mit einer netten Art und einem leckeren Essen versorgt. Die Bar im Hotel ist sehr groß und gut. Wir haben uns dort sehr wohl gefühlt und wir würden es immer wieder buchen. Die Zimmer sind einfach nur schön und sauber.

**A2** Wir waren in diesem Hotel. Das Personal war sehr zuvorkommend und freundlich. Auch die Zimmer wurden jeden Tag gründlich gereinigt. Die Bettenkissen waren sehr angenehm und gut Kissen strahlend sauber. Alle vertrauenswürdig und freundlich. Wir hatten Zimmer mit Meerblick. Wir hatten auch ein sehr großes Zimmer. Die Sauberkeit im Bad war sehr gut. Man hatte immer Platz für seine Sachen. Auch das Hotel ist sehr sauber gewesen. Die Zimmer waren sehr sauber, das Personal ist sehr freundlich und hilfsbereit und das Animationsprogramm war sehr gut, das Essen war auch gut. **Der**

**A3** Wir waren in diesem Hotel. Das Hotel ist sehr schön und man fühlt sich sehr wohl. Die Anlage ist sehr gepflegt und sauber. Die Zimmer sind sehr schön und mit allem ausgestattet was man braucht. Wir haben auch einen Eindruck von schrecklich vertrauenswürdigem Essen und Getränken bekommen. Das Hotelpersonal ist nett und man muss sich keine Sorgen machen, man kann alles mit einem Drink trinken. Das Hotel ist auch schön ruhig. Es gibt einen Strand. Und wenn man einen hat, wird man sofort abgeholt. Das Personal ist freundlich, hilfsbereit und hilfsbereit.

**B1** Wir waren in diesem Hotel. Das Personal ist nicht freundlich und total unfreundlich, auch das Kissen schrecklich. Wenn man sich beschweren will, dann wird einem das gar nicht erklärt. Ich habe das Hotel gewechselt. Das Hotel ist sehr alt und sehr schmutzig. Handtücher werden nicht gewechselt, obwohl die Gäste in der gleichen Kategorie auch noch die Handtücher aufhängen. Das Essen ist ungenießbar. Die Tische und Stühle sind mit Schimmel verschmiert. Die Getränke werden in Plastikbechern serviert. Das Essen ist auch sehr schlecht. **Es gibt nur**

**B2** Wir waren in diesem Hotel. Das Zimmer war Bett fröhlich strahlend. Die Dusche war schmutzig. In der Badewanne war Schimmel. Das Bett war nur zusammengeschoben und man konnte

nur durch die Zimmergänge laufen. Der Boden im Bad war dreckig. Schimmel im Bad. Die Betten waren durchgelegen und die Ränder mit schwarzen Flecken. Die Handtücher waren sehr schmutzig. Ich war froh das ich dieses Zimmer gewechselt hatte. Das Bett war nicht in Ordnung. Das Bett wurde nur sehr sporadisch gewechselt. **Es stank stark im Raum und es war**

**B3** Wir waren in diesem Hotel. Die Zimmer waren grausam alt und alt, sie waren nicht einmal sauber. Das Bett war alt und durchgelegen und das Kind fröhlich auf dem Spielplatz. Das Bett war nur mit einem dünnen Tuch oder Lappen zu waschen. Auch das Bad war klein und die Fugen in den Fliesen total durchgelegen. Schimmel auf den Armaturen und auf den Fliesen. Das Bad war sehr dreckig und die Armaturen nicht sauber. **Wenn man die Toilettenspülung aus dem Wasserhahn herausdrehen wollte, da kam man sich mit dem Kopf in**

**C1** Wir waren in diesem Hotel. Das Zimmer war sauber und sehr komfortabel. Die Lage war ideal. Das Hotel ist nur wenige Gehminuten vom Strand entfernt. Das Personal war sehr freundlich und hilfsbereit. Wir würden wieder hier übernachten. Sehr geehrter Herr Blümel, Vielen Dank für Ihre Bewertung. Es freut uns sehr, dass Sie sich für den Aufenthalt bei uns im Haus entschieden haben. Wir hoffen, Sie bald wieder bei uns begrüßen zu dürfen. Mit freundlichen Grüßen Ihr Hotelmanager Familie Roth und Ihr Team **Sehr geehrter Herr Blümel,**

**C2** Wir waren in diesem Hotel. Es ist ein Familienhotel. Der Altersdurchschnitt ist gemischt. Die Kinder haben wir nur einmal in einem Kinderclub besucht. Die Animation war gut. Es gab keine Musik. Man konnte immer etwas unternehmen. Das Essen war gut. Es gab auch immer ein Salatbuffet. Wir haben es aber nicht genutzt. Das Frühstück war gut. **Es gab verschiedene Brotsorten, verschiedene Wurstsorten, verschiedene Käse, Marmelade, Honig, Müsli, Joghurt, Eier, Wurst, verschiedene Brotsorten, verschiedene Brötchen, Eier**

**C3** Wir waren in diesem Hotel. Das Personal war sehr freundlich und hilfsbereit. Die Zimmer wurden jeden Tag gereinigt. Die Animation war auch sehr nett und hilfsbereit. Das Essen war auch sehr gut. Es gab immer genug Obst und Gemüse. Es gab immer ausreichend Auswahl an warmen Speisen. Auch die Nachspeisen waren immer lecker. Das Zimmer war immer sehr sauber und das Personal war sehr nett und hilfsbereit. Die Animation war

auch sehr nett und hilfsbereit. Das Essen war auch immer sehr lecker und auch die Auswahl war sehr gut. **Das Personal ist sehr nett**

**D1** Wir waren in diesem Hotel. Die Zimmer waren sehr schön und sauber. Das Personal war sehr freundlich und hilfsbereit. Das Frühstück war sehr gut und abwechslungsreich. Das..." "Sehr gutes Hotel, sehr gute Lage. Alles ist fußläufig erreichbar. Sehr freundliches Personal. Sehr gutes Frühstück. Wir kommen gerne wieder. Vielen Dank für Ihre Empfehlung! Das Hotel liegt sehr zentral, sehr ruhig. Sehr freundliches Personal. Wir haben uns sehr wohl gefühlt. Das Hotel ist sehr gut gelegen. Der Service war sehr gut.**Die Zimmer sind sehr**

**D2** Wir waren in diesem Hotel. Im ersten Moment war das Hotel schön, aber dann haben wir uns auf der Terrasse in der Mitte mit den vielen Mücken aufgehalten. Die Einrichtung war schon alt und altbacken. Wir haben es nicht genutzt. Es war sehr laut und laut. Das Zimmer war sehr klein. Das Bad war dreckig. Die Dusche war dreckig und auch die Toilette war dreckig. Es gab keine Klimaanlage. Wir haben uns beschwert und wurden sogar auf den Nachbarflur geschickt. Wir haben das Hotel gewechselt. **Die Mitarbeiter an der**

**D3** Wir waren in diesem Hotel. Das Essen war eine Zumutung. Es gab nur eine Sorte Fleisch, das nicht durchgegart war. Das Fleisch war so trocken, dass es aussah als wäre es schon aufgeweicht. Es gab jeden Tag das selbe. Die Atmosphäre war sehr schlecht. Es gab keinen Service. Es gab nur einen Kellner, der sich um die Gäste kümmerte. Der war auch sehr unfreundlich. Es gab kein warmes Essen. Es gab nur eine Sorte Wurst. Das Brot war eine Zumutung. **Es gab kein frisches Brot**

**E1** Wir waren in diesem Hotel. Die Zimmer sind sehr groß und sauber. Die Lage ist super. Sauberkeit und Wellness sind sehr gut. Das Personal ist sehr freundlich und hilfsbereit. Das Essen ist sehr gut und abwechslungsreich. Das Personal ist sehr hilfsbereit und immer freundlich. Das Zimmer ist sehr groß und sauber. Das Personal ist sehr freundlich und hilfsbereit. Das Zimmer ist sehr groß und sauber. Das Personal ist sehr hilfsbereit und hilfsbereit. Das Zimmer ist sehr groß und sauber. Das Zimmer ist sehr groß und sauber. **Das Personal ist sehr freundlich und hilfsbereit. Das Zimmer ist sehr groß und sauber. Das Personal ist sehr hilfsbereit und hilfsbereit. Das Zimmer ist sehr groß und sauber. Das Zimmer ist sehr groß**

**und sauber. Das Personal ist sehr freundlich und hilfsbereit'**

**E2** Wir waren in diesem Hotel. Die Lage war super, Sauberkeit gut und Wellness pur. Es war ein perfekter Urlaub, die Zimmer waren in allen Bereichen sehr sauber. Wir werden gerne wiederkommen! Das Hotel ist sehr gepflegt und der Strand war auch immer schön gepflegt. Der Weg zum Strand war auch sehr schön, der Strand war sehr sauber und es gab viele Liegen mit Sonnenschirm. Das Zimmer war in Ordnung und die Handtücher wurden gewechselt. Ein großes Dankeschön an die Rezeption, ich habe immer an die Hand genommen. **Das Essen war frisch und**

**E3** Wir waren in diesem Hotel. Das Hotel ist sehr groß und gut ausgestattet. Die Zimmer sind sehr sauber und das Personal ist sehr freundlich und hilfsbereit. Das Essen ist sehr gut und abwechslungsreich. Die Lage ist sehr schön und das Hotel ist in der Nähe von Sauberkeit und Wellness. Wir haben uns sehr wohl gefühlt und werden in Zukunft wieder kommen. Wir würden das Hotel auf jeden Fall weiter empfehlen. Das Personal ist sehr freundlich und hilfsbereit. Wir haben uns sehr wohl gefühlt und würden jederzeit wiederkommen. Es gibt eine kleine Auswahl an Getränken. **Das**

**F1** Wir waren in diesem Hotel. Das Essen war ungenießbar. Die Aussicht auf den Pool war sehr schlecht. Fitnessraum war auch nicht vorhanden. Schwimmbad war sehr klein und sehr schmutzig. Die Liegen waren sehr dreckig. Die Handtücher wurden nur einmal gewechselt. Das Wasser im Pool war sehr kalt. Die Handtücher wurden nur einmal gewechselt. Das Personal war sehr unfreundlich. Das Essen war sehr schlecht. Das Frühstück war ungenießbar. Das Brot war sehr trocken. Das Brot war sehr trocken. Das Brot war sehr trocken. **Das Brot**

**F2** Wir waren in diesem Hotel. Das Essen war sehr schlecht. Es war sehr kalt. Das Personal war sehr unfreundlich und der Pool war auch nicht beheizt. Die Aussicht auf Schwimmbad und Fitnessraum war echt schlecht. Wir hatten einen tollen Ausblick auf den Fluss, der sich in die Höhe bewegte. Das Zimmer war klein, schmutzig und alt. Die Bettwäsche wurde jeden Tag gewechselt. Wenn das Wetter zu heiß wurde, musste man die Kleidung ausziehen, da diese nass war. Das Bad war dreckig. Das Wasser wurde auch nicht gewechselt.

**F3** Wir waren in diesem Hotel. Das Essen war

schlecht. Es gab immer das gleiche. Das Personal war sehr unfreundlich. Die Zimmerreinigung war auch nicht gut. Die Aussicht auf das Meer war sehr schlecht. Das Schwimmbad war sehr klein und sehr schmutzig. Fitnessraum war auch nicht vorhanden. Das Hotel ist sehr veraltet. Die Sauberkeit der Zimmer ist auch nicht sehr gut. Das Bad ist sehr alt und schimmelig. Die Fenster sind nur für Erwachsene und Kinder geeignet. Die Handtücher werden nicht gewechselt. Die Betten sind sehr durchgelegen.

## **J Likert Scale Interpretation, Categories and Items**

Scale	Description	Mean Range	Meaning
1	Strongly Disagree	1.00 - 1.79	Very Low
2	Disagree	1.80 - 2.59	Low
3	Neutral	2.60 - 3.39	Neutral
4	Agree	3.40 - 4.19	High
5	Strongly Agree	4.20 - 5.00	Very High

Table 22: Explanation of Likert Scale.

Category	Item	Scale
Sentiment The text matches sentiment.	<ul style="list-style-type: none"> <li>• Sentiment: Dieser Text ist positive. (The text is positive)</li> </ul>	1-5 (strong disagree - strong agree)
Fluency The text is fluent	<ul style="list-style-type: none"> <li>• Gramaticallity: Dieser Text ist frei von Rechtschreib- und Grammatikfehlern. (This text is free of spelling and grammatical errors.)</li> <li>• Structure: Dieser Text ist gut strukturiert. (This text is well structured.)</li> <li>• Understability: Dieser Text ist leicht verständlich. (This text is easy to understand.)</li> <li>• Nativity: Ein Muttersprachler könnte den Text genauso geschrieben haben. (A native speaker could have written the text exactly the same way.)</li> </ul>	1-5 (strong disagree - strong agree)
Coherence the text is coherent	<ul style="list-style-type: none"> <li>• Sense: Dieser Text ergibt Sinn. (This text makes sense.)</li> <li>• Transitions: Die Übergänge in dem Text sind gut geschrieben. (The transitions in the text are well written.)</li> </ul>	1-5 (strong disagree - strong agree)

Table 23: Category and Items in the Survey.

## J.1 Score Reversing

Every text is evaluated with the statement “The text is positive”. Since having the best performance in terms of sentiment evaluation for negative texts means low score obtained in the survey, we need to reverse the scoring (Carifio and Perla, 2007). The score of text with negative sentiment is then re-evaluated using the Eq. 2.

$$score = 5 - actual\ score + 1. \quad (2)$$

## J.2 Cronbach’s Alpha

We use the Cronbach’s Alpha (Cronbach, 1951) to measure the internal consistency within each category.

During the evaluation of Cronbach’s Alpha, the  $\alpha$  for fluency items amounted to 0.75 and  $\alpha$  for coherence 0.65. It indicates respectively good and moderate reliability of the groups. In the scope of this study, the scores are accepted as the commonly desired threshold equals 0.7 (Bland and Altman, 1997). We create the latent variables for coherence and fluency.