# Synthetic German Product Reviews via LLMs: A Large-Scale, Balanced Dataset for Detecting AI-Generated Opinion Spam

**Johanna Bopst and Melanie Siegel**
Darmstadt University of Applied Sciences
melanie.siegel@h-da.de

## Abstract

The number of manipulated online reviews ('opinion spam') has been increasing for years. This trend is being exacerbated by the widespread use of large language models (LLMs). This makes reliable methods for combatting opinion spam increasingly relevant. While detection methods are advancing, high-quality, publicly available datasets, especially for German, are scarce.

This work introduces SynthReview-DE, a large-scale, balanced dataset of 15,000 German product reviews (50% genuine, 50% LLM-generated) across five categories (books, mobile apps, music, PC, toys). We generate synthetic reviews using four state-of-the-art LLMs (GPT-4.1 Nano, Llama-3.1/3.3, Gemma-3) via few-shot prompting and fine-tuning. Our dataset is carefully constructed to preserve real-world distributions in length, rating, and category. We evaluate linguistic quality using SelfBLEU and demonstrate that both logistic regression and RoBERTa can distinguish synthetic from genuine reviews with high accuracy. The dataset will be publicly released to advance research in automated opinion spam detection.

## 1 Introduction

Online product reviews have a direct influence on purchasing decisions and are a central component of the information search process for many users prior to making a purchase (Sahoo et al., 2025). At the same time, the occurrence of manipulated reviews ("opinion spam") has been increasing for years. This trend is being reinforced by the widespread use of large language models (LLMs), as automatically generated reviews can now be created in large quantities, quickly and cost-effectively. In addition to the loss of trust caused by manipulated reviews (Dwivedi et al., 2021), false reviews also have economic implications. For example, they influence ranking algorithms on platforms and can thus alter the visibility and sales of products (Gobi and Rathinavelu, 2019). This situation increases the relevance of reliable methods against opinion spam. However, there are currently few suitable datasets that can be used to develop and evaluate such classification models, especially for the German language. Available reference datasets are often very limited, for example due to an insufficient number of reviews (Ott et al., 2011). In addition, existing data sets are mostly based on human-written opinion spam, while machine-generated spam has hardly been systematically recorded to date.

This work introduces SynthReview-DE, a high-quality labelled data set of 15,000 German product reviews across five high-traffic categories that includes both genuine and artificially generated texts. It can be used as a basis for research in the field of automated opinion spam detection. We generate synthetic reviews using four state-of-the-art LLMs (GPT-4.1 Nano, Llama-3.1/3.3, Gemma-3) via few-shot prompting and fine-tuning, ensuring realistic distributions in length, rating, and category. Unlike prior work focused on human-written spam, our dataset enables the evaluation of detection models against LLM-generated content — a critical but underexplored frontier in opinion spam research.

Our contributions are:

- The first large-scale, publicly available dataset of LLM-generated German product reviews
- A systematic comparison of generation methods
- Empirical analysis of detection performance across text length and rating levels

In chapter 2, we will provide an overview of the available and comparable datasets. We will then explain the dataset creation process in Chapter 3. In Chapter 4, we will analyse our dataset and conduct classification experiments to demonstrate its

usability for opinion spam classification. Chapter 5 summarises the work and provides an outlook.

## 2 Related Work

Although there are several datasets available for detecting opinion spam, they all have specific limitations concerning our specific task of detecting automatically generated German-language opinion spam. The most frequently used dataset is probably the one by Ott et al. (2011). It comprises 400 genuine reviews of the 20 most popular hotels in a US city and 400 fake reviews of the same 20 hotels, which were created manually by crowd workers recruited via Amazon. The total of 800 reviews is too small to efficiently train today's classification methods and the language is English. Furthermore, only positive reviews with a length of more than 150 characters were considered, which does not correspond to a real distribution (Ott et al., 2011).

In their work, Sandulescu and Ester (2015) created a Yelp dataset in addition to the dataset from Ott et al. (2011), in which all reviews were divided into "recommended" and "not recommended" in advance by the company itself using an internal filter. However, it is not directly clear on what basis this classification is made, for example, whether it is based on the behaviour of the reviewers or on the analysis of the texts (Siegel, 2016). In addition, a data set from Trustpilot was used, in which the reviews had also already been filtered by the company.

Salminen et al. (2022) published a data set balanced in terms of sentence length, rating and product category, consisting of 20,000 English-language reviews generated by GPT-2 and ULMFiT, as well as 20,000 genuine Amazon product reviews.

Siegel (2016) worked with students to investigate opinion spam on the German-language Amazon portal and created a corpus containing several hundred examples. This corpus is too small for the development and evaluation of automatic learning methods and does not contain automatic generated reviews.

Table 1 shows SynthReview-DE and its relation to other datasets.

## 3 Dataset Creation

Figure 1 shows the process of dataset creation.

Both for the genuine and the generated reviews, the data set The Multilingual Amazon Reviews Corpus by Keung et al. (2020) was used. The data was retrieved on 13 December 2017. The data set consists of Amazon product reviews between 1997 and 2015. The reviews originate from a time when LLMs were not yet very widespread. Therefore, it can be ruled out that machine-generated reviews are present in the data set. However, it cannot be ruled out that human-generated opinion spam is included. The raw dataset comprises approximately 679,000 reviews.

The first step was preprocessing of the data, such as deletion of incomplete, English-language and duplicated entries, as well as special characters, such as HTML code. Based on the cleaned data, a dataset of real and generated reviews was created. From the total of 32 categories in the data set, we selected five high-traffic categories (books, mobile apps, music, PC, toys) to ensure sufficient review volume and diversity in product types. Stratified sampling was applied to preserve the original distribution of star ratings (1-5) and word counts across categories. For each category, we extracted 5,000 reviews, ensuring balanced representation. This ensures sufficient data per category for robust model training and evaluation, while remaining computationally feasible.

### 3.1 Few-Shot Prompting

Review generation using few-shot prompting was implemented using the Python library OpenAI (OpenAI, 2025). Each of the models (GPT-4.1 Nano, Llama-3.1-SauerkrautLM-70b-Instruct, Llama-3.3-70B-Instruct and Gemma-3-27b-It) generated 300 reviews per category.

Five genuine reviews were selected as examples from the pre-processed data set for each category using a fixed seed. Since this study generates five reviews with up to five stars, k = 5 is selected, so that one example review is used for each rating. The product name and star rating served as input, while the content of the review was specified as output. The user prompt, which represents the actual instruction to the model, contained the product name, star rating, and a required approximate word count, analogous to the few-shot examples. For this purpose, the corresponding attribute values of the genuine reviews were also randomly taken from the pre-processed dataset. To avoid excessive pattern formation, a separate seed was used for each model so that the models did not generate reviews for the same products. In this way, all four models received identical few-shot examples,

| Dataset | Language | Type | Size | Public? |
|---|---|---|---|---|
| Ott et al. (2011) | English | human-written | 800 | yes |
| Sandulescu and Ester (2015) | English | human-written | 67,000 | no |
| Salminen et al. (2022) | English | LLM-generated | 40,000 | yes |
| Siegel (2016) | German | human-written | 280 | yes |
| **SynthReview-DE (this work)** | **German** | **LLM-generated** | **15,000** | **yes** |

Table 1: Datasets in the context of opinion spam detection

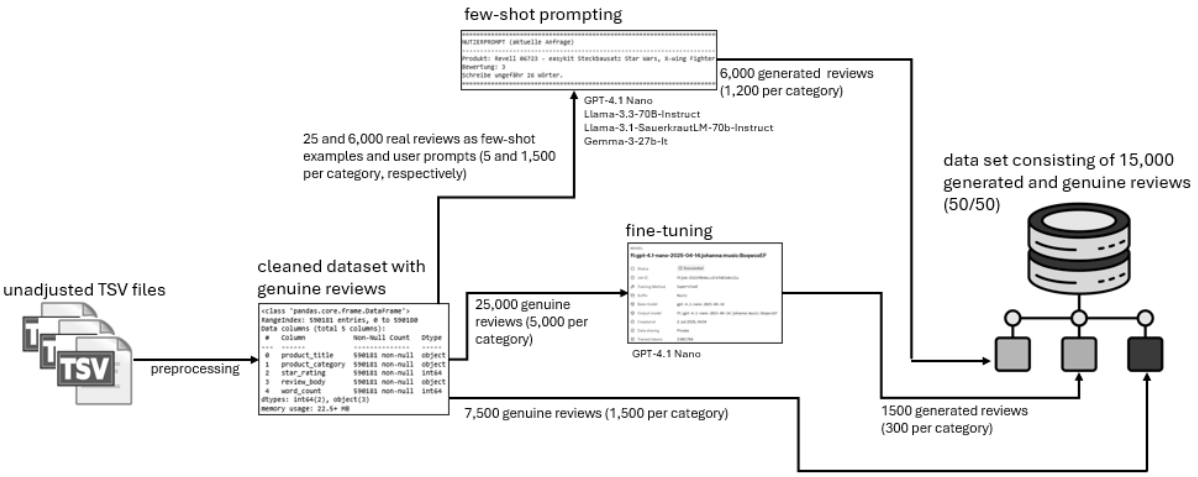


Figure 1: Process of dataset creation

but different user prompts for the 1,500 reviews to be generated. To increase the creativity of the generated reviews, the top-p sampling decoding strategy with a p-value of 0.9 was used. It offers a dynamic alternative to deterministic methods such as beam search and rigid top-k sampling. Appendix A shows an example for a generation prompt.

### 3.2 Generation Using Fine-Tuning

In addition to few-shot prompting, the GPT 4.1-Nano model was also used for fine-tuning. The functionality provided by OpenAI was used for this purpose. A separate model was fine-tuned for each of the five selected categories: books, mobile apps, music, PC and toys. This was based on a JSON file prepared for each category, which contained the corresponding training data. Each entry in the file consisted of a standardised message structure analogous to the structure in few-shot prompting. The models were fine-tuned using the standard hyperparameters recommended by OpenAI. The number of epochs was set to three. After each adjustment, the training loss and train mean token accuracy were calculated. Choosing a small batch size results in the model parameters being updated more frequently. This promotes more stable convergence, reduces the risk of local overfitting and, at the same time, leads to denser and noisier curves. Another key hyperparameter was the learning rate multiplier, which was set to 0.1. Reducing the original learning rate results in smaller adjustment steps. This results in a slower but more stable approach to the target distributions of the training data. Figure 2 shows the training loss curve for fine-tuning the GPT model for the music category. The model's learning curve shows a steady and consistent decline from approximately 3.0 to 2.45, indicating stable learning progress. After approximately 1,100 training steps, the decline in loss decreases, indicating that the model has largely converged. There is no renewed increase, which would indicate overfitting.

In addition, an initial attempt was made to fine-tune the Mistral-7B model using LoRA. The generated reviews were semantically less coherent, contained more repetitive sequences and appeared less realistic overall. Due to these unsatisfactory results, the approach was rejected and not pursued further.

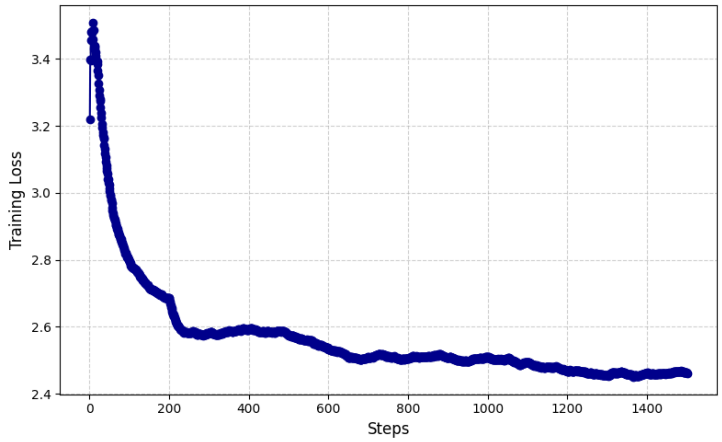


Figure 2: Loss curve for fine-tuning the LLM on reviews of the category music

| Model | Number of Reviews |
|---|---|
| GPT-4.1 Nano | 1.500 |
| GPT-4.1 Nano (fine-tuned) | 1.500 |
| Lama-3.1-SauerkrautLM-70b-Instruct | 1.500 |
| Llama-3.3-70B-Instruct | 1.500 |
| Gemma-3-27b-It | 1.500 |
| Genuine reviews | 7.500 |

Table 2: Number of generated reviews per model

## 4 Dataset Analysis and Classification Experiments

The created dataset comprises a total of 15,000 reviews, of which 50% are genuine texts and 50% are generated texts. We therefore have an artificially balanced dataset that enables machine learning. For the generation, the four different LLMs GPT-4.1 Nano, Llama-3.1-SauerkrautLM-70b-Instruct, Llama-3.3-70B-Instruct and Gemma-3-27b-It were used. While all models were used with few-shot prompting, the GPT-4.1 Nano model was additionally fine-tuned. Each model contributed 1,500 reviews to the dataset, ensuring a balanced distribution of sources (see Table 2). In addition, the number of categories is evenly distributed. Each of the five categories contains 3,000 reviews. This is relevant in order to ensure comparability when classifying the generated reviews.

### 4.1 Analysis

A key objective was to ensure that the genuine and generated reviews differed as little as possible in their basic distribution. This applies both to the selection of genuine reviews from the original dataset of genuine Amazon reviews and to the reviews that were used as prompts for the models. In the prompts, the word count, star rating and title were provided in each case, so that the generated reviews have the same structural characteristics as the genuine ones.

In addition to statistical analysis on distributions of text length, star rating and category, it is crucial for the evaluation of the data set to examine the linguistic quality of the reviews. This involves analysing the extent to which genuine and generated texts differ in terms of their diversity. Self-BLEU was used to evaluate the linguistic diversity of the reviews. The metric calculates the similarity of a review to reviews from a randomly selected subset (Zhu et al., 2018). Lower values indicate greater linguistic diversity, while higher values indicate more homogeneous and repetitive structures. For the calculation, a total of 1,000 reviews were randomly selected and compared with 200 reference reviews each. The average value was then calculated. Table 3 shows the BLEU and N-Gram linguistic diversity of the generated and genuine reviews. BLEU-n refers to the n-gram diversity of the texts. BLEU-1, for example, measures the agreement of individual words between the texts. BLEU-2 is the bigram diversity, i.e. the proportion of word pairs that also occur in at least one

other text. BLEU-3 and -4 are then trigrams and four-grams, respectively. The overall BLEU score combines these n-gram diversity values using a geometric mean and a brevity penalty. The results in Table 3 show that the genuine reviews exhibit the greatest diversity. The fine-tuned GPT model (GPT-4.1 Nano) achieves the lowest value among the models at 1.26, thus generating the most diverse texts. In contrast, the reviews from the GPT-4.1 Nano (few-shot prompting) model with 1.83 and Llama-3.3-70B-Instruct with 1.76 have the highest values, indicating less variance in the wording.

The group of combined models performs second best, which suggests that the different language models prefer different words and phrases when generating text and that their linguistic patterns complement each other. It is also striking that the two GPT models use the same words most frequently (BLEU-1). The non-fine-tuned GPT-4.1 Nano and Llama-3.3-70B-Instruct models use the same phrases most frequently (BLEU-2 - BLEU-4). Appendix B shows some examples of generated and genuine reviews in the category *Music*.

### 4.2 Classification Experiments

After examining the quality of the data set in the previous section, we will now examine the extent to which the generated reviews can be automatically distinguished from the real ones. The aim is to analyse how easily the generated reviews from two different models can be classified as such. Logistic regression and RoBERTa were used as classifiers. The comparison between a linear model and a transformer model enables an informed assessment of the extent to which more complex linguistic patterns contribute to the detection of generated reviews.

For classification using logistic regression, the review texts were first converted into a numerical representation. For this purpose, TF-IDF vectorisation was used, in which each word in the corpus is assigned a weight. Logistic regression was then trained on these vectors. The cross-entropy loss function is minimised using the SGD method. To avoid overfitting, L2 regularisation is incorporated. Suitable hyperparameters (number of words to be considered in TF-IDF and number of iterations in logistic regression) were selected using a grid search in combination with 5-fold cross-validation. The training set was divided into several subsets so that the model is repeatedly trained and evaluated on different combinations of training and validation data. This allows the generalisability of the model to be better assessed and the risk of overfitting to be reduced.

In logistic regression, each word in the vocabulary is assigned a coefficient that expresses its significance for classification. Negative coefficients indicate that the occurrence of a word is more frequently associated with authentic reviews, while positive coefficients indicate that a word is more likely to occur in generated reviews. Table 4 shows the most relevant words for classification.

Analysis of the logistic regression coefficients reveals clear linguistic differences that are decisive for assignment to the respective classes. Words with negative coefficients are predominantly function words such as 'the', 'but', 'as' or 'in'. These indicate that authentic reviews are written in a more everyday, unobtrusive language that contains typical grammatical structures and nuances. Terms such as 'not' or 'already' also indicate a differentiated style of expression that is characteristic of naturally written texts. In contrast, fake reviews have strongly positive coefficients for mostly positive evaluative words such as "perfekt" (perfect), "sehr" (very) or "empfehlenswert" (recommendable). These indicate exaggerated language that is less common in genuine reviews. In addition, the pronoun "ich" (I) occurs more frequently in fake reviews, suggesting that LLMs attempt to simulate authenticity by emphasising the first-person perspective. The fact that words with high negative coefficients are predominantly capitalised even though in German they are only capitalised at the start of a sentence, which also indicates that the generated reviews often begin with similar phrasing.

Secondly, RoBERTa (Liu et al., 2019) was used for classification. For the present application, the model was adapted to the binary classification task in a fine-tuning process. The trainable parameters were updated based on the training data to optimise the model's ability to distinguish between genuine and generated reviews.

The results of the classification on the test data show that both models are able to reliably distinguish between genuine and generated reviews. Logistic regression achieves an accuracy of 88.3% with an F1 score of 0.88. RoBERTa achieves slightly better overall results with an accuracy of 90.1% and an F1 score of 0.91.

Table 5 shows the results of the classification experiments. It is striking that RoBERTa achieves a

| Model | BLEU | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|---|
| GPT-4.1 Nano | 1.83 | 26.91 | 4.00 | 1.59 | 0.81 |
| GPT-4.1 Nano (fine-tuned) | 1.26 | 24.22 | 2.38 | 0.95 | 0.49 |
| Lama-3.1-SauerkrautLM-70b-Instruct | 1.47 | 24.09 | 3.27 | 1.20 | 0.63 |
| Llama-3.3-70B-Instruct | 1.76 | 24.12 | 3.81 | 1.63 | 0.86 |
| Gemma-3-27b-It | 1.52 | 23.86 | 3.20 | 1.25 | 0.66 |
| all models | 1.35 | 22.96 | 2.86 | 1.12 | 0.58 |
| genuine reviews | **0.91** | **17.65** | **1.65** | **0.75** | **0.38** |

Table 3: BLEU and N-Gram linguistic diversity of generated and genuine reviews

| **Genuine Reviews** | | **Generated Reviews** | |
|---|---|---|---|
| **Word** | **Coefficient** | **Word** | **Coefficient** |
| der (the, masc) | -2.87 | Die (the, fem) | 8,42 |
| doch (but) | -2.13 | Ein (a) | 5,60 |
| als (as) | -2.12 | Sehr (very) | 4,97 |
| nun (now) | -2.08 | Klare (clear) | 4,62 |
| diesen (this) | -2.08 | okay (ok) | 3,98 |
| am (on) | -2.01 | Ich (I) | 3,78 |
| bzw (or) | -2.00 | perfekt (perfect) | 3,75 |
| nicht (not) | -1.99 | empfehlenswert (recommendable) | 3,74 |
| schon (already) | -1.86 | aber (but) | 3,68 |
| im (in) | -1.85 | Perfekt (perfect) | 3,62 |

Table 4: Top 10 words by coefficients for logistic regression

recall of 95.5%, which is substantially higher than that of logistic regression (86.3%). This indicates that RoBERTa is significantly better at correctly identifying fake reviews and thus misses considerably fewer fake instances. Logistic regression, on the other hand, achieves a slightly higher precision of 89.9% compared to RoBERTa (86.2%). This suggests that logistic regression classifies fewer genuine reviews as fake and therefore produces fewer false positive predictions.

Overall, while logistic regression is more conservative in labeling reviews as fake, RoBERTa demonstrates superior performance in terms of recall and F1-score, indicating a better overall balance between precision and recall when detecting fake reviews.

To investigate whether the two models systematically misclassify the same reviews, the misclassifications of RoBERTa and logistic regression were compared. RoBERTa and logistic regression exhibit different types of errors, and their misclassifications overlap only to a small extent. In the generated reviews, misclassifications rarely occur simultaneously. This indicates that RoBERTa is more robust and makes only isolated errors in classification (except for GPT-4.1 Nano (finetuned)).

These results suggest that RoBERTa makes greater use of contextual patterns and sentence structures, while logistic regression primarily responds to word distributions. The models thus complement each other to some extent in their error patterns. Looking at misclassifications and correct classifications together, the models make the same decision in 85.1% of cases. This could be an indication that an ensemble approach could potentially further improve overall performance.

The analysis of classification accuracy as a function of text length also highlights differences between the two classifiers examined. For short reviews with fewer than 50 words, both RoBERTa (89.7%) and logistic regression (86.9%) achieve comparatively moderate levels of accuracy.

In the medium length range between 50 and 200 words, the performance of both models increases significantly. Longer texts in this segment prove to be particularly advantageous for RoBERTa, as the diversity of lexical and syntactic structures they contain optimally supports the context sensitivity of transformer-based models. However, it is noticeable that logistic regression achieves temporarily better results than RoBERTa for texts in the range of 150 to 200 words, with 93.4%. It is possible that

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.883 | 0.899 | 0.863 | 0.881 |
| RoBERTa (Epoch 1) | 0.901 | 0.862 | 0.955 | 0.906 |

Table 5: Performance of both classifiers

linear models can benefit from the direct weighting of characteristic keywords in clearly structured, medium-sized texts.

The investigation of accuracy as a function of star ratings shows that both models tend to perform better with higher ratings (see Figure 3). The biggest difference in accuracy exists in the classification of one-star reviews, where logistic regression performs significantly worse. This suggests that linear models are more dependent on unambiguous keywords and phrases, which may occur less frequently or more variably in very negative reviews.

It is also striking that both RoBERTa and logistic regression achieve lower accuracy for one- and five-star reviews than for four-star reviews. One possible explanation for this lies in the linguistic sophistication of four-star reviews. These often contain a mixture of praise and mild criticism, resulting in linguistic patterns that are more suitable for identifying generated reviews.

The experiments demonstrate that the dataset is suitable for developing classification models with decent performance.

## 5 Conclusions

We presented SynthReview-DE, the first large-scale, balanced dataset that includes both genuine and LLM-generated reviews. Synthetic reviews were generated using few-shot prompting of various models and targeted fine-tuning. For few-shot prompting, five genuine reviews per category served as examples, which were provided to the models as input. The prompts also contained structuring information such as word count, star rating, and product title, which helped the models to generate coherent and realistic reviews. This enabled the generation of consistent and realistic texts without having to adjust the model parameters themselves. The fine-tuning involved directly adjusting the model weights to the specific style and word choice of German reviews. The selection of training and prompt examples from the real reviews was done using stratified sampling to ensure representativeness in terms of category, star rating and word length. The descriptive analysis showed that the distributions of star ratings and text lengths reflect the original data well. Overall, the data set is therefore statistically consistent and suitable for extensive analysis. Linguistically, the generated texts showed less diversity than the real reviews. Our results show that fine-tuned Models produce more stylistically diverse, realistic reviews.

The classification attempts with RoBERTa and logistic regression yielded consistent results. Overall, RoBERTa achieved better classification of genuine and generated reviews. The choice of the generation model has a significant impact on classification. The fine-tuned GPT-4.1 Nano model produced the best quality and most diverse reviews. Furthermore, the classification quality varies systematically with text length and star rating. Very short texts are more difficult to classify due to their low information content, while medium-length texts provide more reliable results. Four-star reviews were recognised most reliably by both models. The work also has some limitations. In the broader context of opinion spam, it must be taken into account that fake reviews appear on platforms in various forms. For example, there are cases in which buyers actually purchase products, rate them positively, and then receive a refund from the companies. Such reviews are formally verified, but still represent a form of manipulated content. A veritable market for this has developed, particularly in the Amazon environment (Hill, 2022). We work with an artificially balanced dataset that enables machine learning. In a real-world environment, however, the data will be highly unbalanced, which could lead to overfitting, for example.

Despite these limitations, compared to existing work, SynthReview-DE closes an important research gap in opinion spam detection, particularly in the German-language domain. While previous datasets often only comprise a few hundred to a thousand reviews or are not publicly accessible, this dataset represents a high-quality basis for further research into the detection of opinion spam. SynthReview-DE enables research into generalizable detection models, adversarial training, and

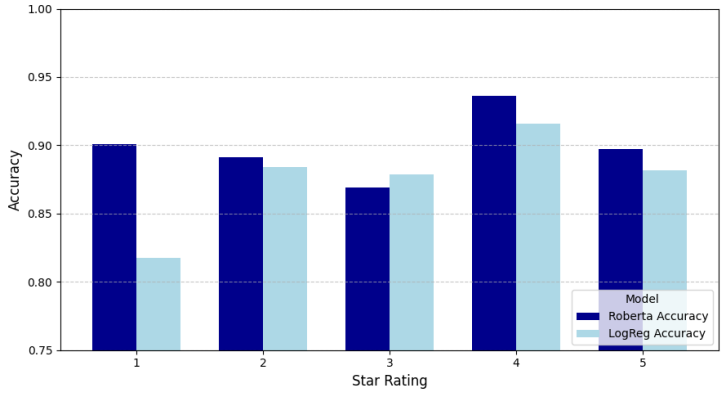


Figure 3: Accuracy of the models in relation to the stars

multi-lingual transfer learning.

The dataset will be released under a research-only license with the publication of this paper to promote ethical, reproducible work.

### Ethical Considerations

While this dataset is intended for research on detecting AI-generated opinion spam, we emphasize that it should not be used to generate more deceptive content. We encourage responsible use, including transparency in model training and evaluation. The dataset will be released under a research-only license to prevent misuse.

## References

Yogesh K. Dwivedi, Elvira Ismagilova, D. Laurie Hughes, Jamie Carlson, Raffaele Filieri, Jenna Jacobson, Varsha Jain, Heikki Karjaluoto, Hajer Kefi, Anjala S. Krishen, Vikram Kumar, Mohammad M. Rahman, Ramakrishnan Raman, Philipp A. Rauschnabel, Jennifer Rowley, Jari Salo, Gina A. Tran, and Yichuan Wang. 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59:102168.

N. Gobi and A. Rathinavelu. 2019. Analyzing cloud based reviews for product ranking using feature based clustering algorithm. *Cluster Computing*, 22(Suppl 3):6977–6984.

Simon Hill. 2022. Inside the underground market for fake amazon reviews. Accessed 14th November 2025.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. *Preprint*, arXiv:2010.02573.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

OpenAI. 2025. Chat completions | openai api reference. OpenAI Platform Documentation. Accessed 11th Dezember 2025.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon gyo Jung, and Bernard J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Vlad Sandulescu and Martin Ester. 2015. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 971–976. ACM.

Melanie Siegel. 2016. Opinion Spam – Meinungsäußerungen als Fake. *Information - Wissenschaft & Praxis*, 67(5-6):304–310.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

## A Example Generation Prompt

The following is an example for a few-shot generation prompt in the category *Toys*:

```
[
  {
    "role": "system",
    "content": "Du bist ein Käufer oder eine Käuferin,
    der/die ehrliche, natürlich klingende Amazon-Produktrezensionen schreibt.
    Vermeide unnötige Wiederholungen.
    Schreibe flüssig und in Alltagssprache.
    Halte dich an den angegebenen Wortumfang.
    Keine formelle Sprache - schreibe so, wie echte Kund:innen schreiben würden."
  },
  {
    "role": "user",
    "content": "Produkt: Janod 4506520 - Maxi Küche Chic, 8 Accessoires
    Bewertung: 1"
  },
  {
    "role": "assistant",
    "content": "Unsere Tochter sollte eine Holz-Kinderküche zu Weihnachten bekommen.
    ... ansonsten können wir nur vom Kauf abraten."
  },
  {
    "role": "user",
    "content": "Produkt: Lego Star Wars 7913 - Clone Trooper Battle Pack
    Bewertung: 2"
  },
  {
    "role": "assistant",
    "content": "Mein Sohn freute sich sehr über das Teil, jedoch mussten wir feststellen,
    dass 4 Klein-Teile fehlten. ... Nicht gut!"
  },
  {
    "role": "user",
    "content": "Produkt: Lego Star Wars 9494 - Anakins Jedi Interceptor
    Bewertung: 3"
  },
  {
    "role": "assistant",
    "content": "Leider zählt man hier den Star-Wars-Zuschlag zu sehr: wenig Teile für viel Geld."
  },
  {
    "role": "user",
    "content": "Produkt: Lego Technic 42023 - Baustellen-Set
    Bewertung: 4"
  },
  {
    "role": "assistant",
    "content": "3 Baustellenfahrzeuge in einem Set ist einfach nur Spitze.
    Freue mich schon auf die großen Augen meines Enkels zu Weihnachten."
  },
  {
    "role": "user",
    "content": "Produkt: AMIGO 07900 - Bohnanza, Kartenspiel
    Bewertung: 5"
  },
  {
    "role": "assistant",
    "content": "Ein Spiel, das ohne viel Material spannend und abwechslungsreich ist.
    Selbst nach mehreren Runden bleibt der Spaß erhalten.
    Glück und Taktik sind gut ausbalanciert."
  },
  {
    "role": "user",
    "content": "Produkt: Revell 06723 - easykit Steckbausatz Star Wars, X-wing Fighter
    Bewertung: 3
    Schreibe ungefähr 26 Wörter."
  }
]
```

## B Example Reviews in the Category *Music*

| Model | Example Review |
|---|---|
| GPT-4.1 Nano | Das Album ist okay, aber nichts wirklich Neues. Manche Songs gefallen mir, andere sind eher durchschnittlich. Für Fans sicher ganz in Ordnung, mehr aber auch nicht. |
| GPT-4.1 Nano (fine-tuned) | U2 hat den alten Sound wieder entdeckt und mit ein wenig neuen Sounds gemixt. Leider keine CD für's Auto, sondern nur was für zu Hause. |
| Llama-3.1-SauerkrautLM-70b-Instruct | Schade, dass es nicht mehr so wie die alten Meisterwerke geworden ist. Noch ok, aber kein Highlight. |
| Llama-3.3-70B-Instruct | Gutes Album, aber nicht überragend. Einige Tracks sind sehr schön, andere eher durchschnittlich. Für Fans von Michael Jackson sicherlich eine Empfehlung, aber nicht sein Meisterwerk. |
| Gemma-3-27b-It | Interessantes Konzeptalbum, aber nicht ganz meins. Die Kombination aus elektronischen Klängen und klassischen Instrumenten ist gewöhnungsbedürftig. Einige Tracks gefallen mir gut, andere wirken etwas zu experimentell und gehen unter. Insgesamt solide, aber kein Highlight. |
| Genuine | War für mich die Bestätigung, dass sich der Musikgeschmack mit dem Alter wirklich verändert. Hatte die Scheibe viel besser in Erinnerung. Höre die neueren Platten aber immer noch gerne... |

Table 6: 3-star example reviews from the music category in the dataset.