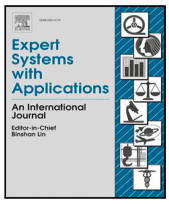
Expert Systems With Applications 261 (2025) 125514

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings

Nils Constantin Hellwig *, Jakob Fehle, Christian Wolff

Media Informatics Group, University of Regensburg, Regensburg, 93040, Bavaria, Germany

ARTICLE INFO

Keywords:
Natural language processing (NLP)
Sentiment analysis (SA)
Aspect-based sentiment analysis (ABSA)
Large language models (LLMs)
Synthetic data generation
Low-resource settings
Data augmentation

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained task in sentiment analysis, aiming to identify sentiment expressed towards specific aspects of an entity. This paper explores the use of Large Language Models (LLMs), specifically GPT-3.5-turbo and Llama-3-70B, for generating annotated data in Aspect-Based Sentiment Analysis (ABSA), aiming to address the scarcity of labelled datasets in the field. Two low-resource scenarios are considered, with 25 and 500 manually annotated examples available. In the 25-example scenario, adding synthetic examples generated through few-shot prompting resulted in F1 scores of 81.33 for Aspect Category Detection (ACD) and 71.71 for Aspect Category Sentiment Analysis (ACSA). For the 500-example scenario, synthetic data augmentation showed a notable gain only for the ACSA task, raising the F1 score from 84.54 to 86.70.

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) is a subfield of sentiment analysis (SA), aiming to provide more nuanced and detailed insights into opinions expressed in textual data. Unlike traditional sentiment analysis, which classifies an entire document (Hellwig, Bink, Schmidt, Fehle, & Wolff, 2023; Tripathy, Anand, & Rath, 2017) or sentence (Liu, 2010) as positive, negative, or neutral, ABSA focuses on identifying sentiments associated with specific aspects or features of a product, service, or entity. This granular approach allows for a deeper understanding of people's opinions, enabling more precise insights (Liu, Hu, & Cheng, 2005; Pontiki, Galanis, Papageorgiou, Manandhar, & Androutsopoulos, 2015).

Similar to other areas of natural language processing research, the field of ABSA faces a lack of annotated corpora across various languages and domains for training machine learning models to recognize aspects and associated sentiment polarities in texts (Chebolu, Dernoncourt, Lipka, & Solorio, 2022; Fehle, Schmidt, & Wolff, 2021). The process of manually annotating corpora is very time-consuming, primarily due to the multi-layered nature of the aspect-based sentiment annotation process, which involves multiple steps: identifying related aspect and sentiment phrases and then categorizing them into aspect and polarity classes (Fehle, Münster, Schmidt, & Wolff, 2023; Li, Wang, Ding, Zhou, & Yan, 2023).

Large language models (LLMs) appear to be a promising approach in this context. LLMs are based on a transformer architecture and are characterized by their large size with billions of parameters. They have shown a comprehensive language understanding and the ability to produce text that is difficult for humans to distinguish from authentic human text in a wide range of language tasks (Brown et al., 2020; Floridi & Chiriatti, 2020). Across document-, sentence- and aspect-level, LLMs have demonstrated commendable zero-, one- and few-shot sentiment classification performance without resource-intensive data annotations (Møller, Dalsgaard, Pera, & Aiello, 2023; Zhang, Deng, Liu, Pan, & Bing, 2023).

In certain situations, a scarcity of domain-specific examples for annotation and training may occur. Moreover, classifying text through a commercial API of proprietary LLMs like GPT-3 may not be allowed due to inherent privacy considerations (Møller et al., 2023). Furthermore, training and inference of LLMs is highly computationally intensive compared to Small Language Models[1] (SLMs) like those based on BERT (Bidirectional Encoder Representations from Transformers) architecture fine-tuned on annotated datasets (Devlin, 2018; Wang, Liu, Xu, Zhu, & Zeng, 2021). To overcome the limitations from data scarcity as well as from computational demands, studies have generated training examples utilizing LLMs both in the broader field of NLP and specifically in the domain of sentence-level SA (Meyer, Elsweiler, Ludwig,

* Corresponding author.
E-mail addresses: Nils-Constantin.Hellwig@ur.de (N.C. Hellwig), Jakob.Fehle@ur.de (J. Fehle), Christian.Wolff@ur.de (C. Wolff).

[1] No generally accepted definition exists for categorizing language models as small or large. In this study, models with less than 1 billion parameters are categorized as small, otherwise as large.

https://doi.org/10.1016/j.eswa.2024.125514
Received 20 August 2024; Received in revised form 29 September 2024; Accepted 4 October 2024
Available online 17 October 2024

Fernandez-Pichel, & Losada, 2022; Møller et al., 2023; Van Nooten & Daelemans, 2023).

This study utilized LLMs to generate annotated examples for ABSA in low-resource scenarios, assuming only 500 or 25 manually annotated (real) examples. LLMs, specifically Llama-3-70B and GPT-3.5-turbo, were used to generate examples. Using increasing numbers of synthetic examples in addition to a fixed set of the given 500 or 25 annotated real examples, SLM were trained for four common ABSA tasks. The performance of the models was compared against that achieved when exclusively using 500, 1,000 or 2,000 real examples for training. Finally, the quality of annotations generated by LLMs was assessed and compared using 2,400 human annotations of synthetic examples.

For a given set of training examples, we fine-tuned state-of-the-art (SOTA) transformer models. Specifically, we considered those emphasized in the comprehensive literature review on ABSA tasks by Zhang, Li, Deng, Bing, and Lam (2022). The following ABSA tasks were considered: (1) Aspect Category Detection (ACD), (2) Aspect Category Sentiment Analysis (ACSA), (3) End-to-End ABSA (E2E-ABSA) and (4) Target Aspect Sentiment Detection (TASD). Finally, a total of 2,400 human annotations of synthetic examples were employed to assess the quality of annotations generated by LLMs.

Supplementary materials, including code and model training results, are provided in the appendix of this work and are accessible on GitHub[2].

2. Related work

2.1. Performance of LLMs in Sentiment Analysis (SA)

In the field of Sentiment Analysis (SA), Large Language Models (LLMs) have showcased commendable performance. Qin et al. (2023), for instance, reported an accuracy of 88.8 in the evaluation of 872 samples from the SST2 (Stanford Sentiment Treebank v2) dataset (Socher et al., 2013) using GPT-3.5 (text-davinci-003) under zero-shot learning conditions. Each prompt consisted of a concise task description, along with the sentence to be classified, to which labels, `POSITIVE` or `NEGATIVE`, were to be assigned (Qin et al., 2023).

Zhang et al. (2023) conducted evaluations on various ABSA tasks, with the prompt including a description of each task along with a list of the considered sentiment polarities. They assessed the performance of determining the sentiment polarity expressed towards a given aspect term in a text. Employing ChatGPT, the evaluation revealed an F1 score of 76.80 for Laptop14 and 82.80 for Rest14. Notably, these datasets were introduced in the SemEval ABSA challenge 2014 (Pontiki et al., 2014) and consist of laptop and restaurant reviews, respectively. Moreover, they assessed ChatGPT's performance in E2E-ABSA, where it was required to identify all aspect terms in a given sentence, along with the sentiments expressed towards them. When evaluating the restaurant domain of the SemEval ABSA datasets from 2014 to 2016, the obtained F1 scores were 54.46, 40.03, and 75.80 for Rest14, Rest15, and Rest16, respectively.

Moreover, Zhang et al. (2023) investigated the impact of utilizing 10 annotated examples per sentiment polarity as few-shot examples. In the case of Rest14 and Rest15, incorporating these few-shot examples improved performance and resulted in an F1 score of 63.30 (+8.95) and 52.85 (+12.82), respectively. However, for Rest16, the performance declined to 59.22 (−16.58).

To further improve on the capabilities of LLMs, Simmering and Huoviala (2023) compared the performance of GPT-4 and GPT-3.5 for the E2E-ABSA task on the Rest14 dataset by investigating few-shot learning and fine-tuning of LLMs. They discovered that the models sometimes encountered difficulties in adapting to the detailed requirements of the ABSA task during few-shot learning, whereas fine-tuned models were able to achieve a better adaptation to the complex properties of the E2E task, resulting in a F1 score of 83.8.

[2] Resources for the paper: https://github.com/NilsHellwig/exploring-absa-llm-augmentation

2.2. Data augmentation to mitigate data scarcity in ABSA

Data augmentation techniques aim to create new training examples by transforming existing data in a constrained way to improve the model's generalization ability (Li et al., 2023). These techniques are particularly valuable when the labelled data is limited or unbalanced (Bayer, Kaufhold, & Reuter, 2022).

In the field of ABSA, back-translation has been employed as a data augmentation approach (Liesting, Frasincar, & Truşcă, 2021; Wang, Jiang, Ma, Liu, & Okazaki, 2023). As an example, Liesting et al. (2021) applied a back-translation technique by translating English sentences into Japanese and subsequently re-translating them back into English. Adding these transformed sentences to the training dataset led to an improvement in accuracy by +0.5 (84.4) when evaluating Rest16. The input of the trained model consisted of both the sentence and a category, for which the polarity had to be identified (Liesting et al., 2021). Furthermore, Liesting et al. (2021) applied random synonym replacements using the WordNet lexicon (Fellbaum, 2010), along with word insertions, deletions, and swaps. Their approach led to an increase of +0.5 (84.4) in accuracy when evaluating on the sentiment polarity detection subtask of Rest16 (Liesting et al., 2021). Similarly, evaluation on the same subtask of Rest15 resulted in an improvement of +1.0 (78.9) in terms of accuracy.

Generative approaches using language models to generate synthetic training examples have been used less frequently to date. In the broader field of SA, Møller et al. (2023) operated under the assumption of a low-resource scenario, utilizing a dataset comprising 500 real examples from the SemEval-2017 Task 4: "Sentiment Analysis in Twitter" (Rosenthal, Farra, & Nakov, 2017). A prompt was executed for each individual real example from the dataset. LLMs received the example (one-shot learning) and its associated sentiment as inputs to generate 10 examples simultaneously. Subsequently, an investigation with both ChatGPT and GPT-4 was conducted to ascertain whether using 500 synthetic examples, coupled with a gradual increase in size up to 5000 samples, could yield an improved classification performance compared to training based solely on real examples. Their findings revealed that synthetic data exhibited a lower classification performance than human-annotated data. However, it should be noted that a direct comparison was made between exclusively synthetic examples on one hand and exclusively real sentences on the other hand. The synthetic examples were thus not added to the existing real sentences for training.

In the field of ABSA, Li, Chen, Quan, Ling, and Song (2020) utilized a masked sequence-to-sequence method for data augmentation in order to tackle the ATE task. They fine-tuned a MASS (MAsked Sequence to Sequence pre-training) model (Song, Tan, Qin, Lu, & Liu, 2019) in order to generate new sentences by masking half of the tokens in a sentence while preserving the text and position of the aspect terms. For each example of the dataset of real examples, one synthetic example was generated using the MASS model. The performance of ATE was improved by adding synthetic examples to real examples when evaluating a BERT model used for token classification in order to determine the aspect terms, as proposed by Xu, Liu, Shu, and Yu (2019). Evaluating the ATE subtask of Rest16 revealed an F1 score of 80.29, which was an increase of +2.8 compared to the setting without augmented examples. Similarly, the evaluation of the ATE subtask of Laptop14 revealed an improved F1 score when augmented examples were included, rising from 84.59 to 85.33 (+0.74).

3. Methodology

3.1. Study design

In this study, we considered two low-resource scenarios. Following the example of low-resource scenarios studied by Møller et al. (2023) and Zhang et al. (2023), the first low-resource scenario assumed the

Table 1
Experimental conditions.

Conditions	C1	C2	C3	C4	C5	C6	C7	C8	C9
# Real Examples	500	1,000	2,000	500	500	500	25	25	25
# Synth Examples for LRS_{500}	0	0	0	500	1,000	1,500	0	0	0
# Synth Examples for LRS_{25}	0	0	0	0	0	0	475	975	1,975
# Total Examples	500	1,000	2,000	1,000	1,500	2,000	500	1,000	2,000

availability of 500 annotated sentences (subsequently referred to as LRS_{500}). For these examples, it was assumed that annotations exist for all aspect terms, their corresponding aspect categories and the sentiment expressed towards each aspect term. An LLM was utilized to generate 500, 1,000 or 1,500 examples using 25 few-shot examples randomly selected from the 500 real annotated sentences.

The incorporation of synthetic examples was evaluated for various ABSA tasks. We assessed whether incorporating synthetic examples along the 500 real examples enhances performance compared to relying solely on the 500 real examples. Additionally, performance was compared against conditions involving only 1,000 or 2,000 real examples, resulting in a total of six conditions (C1–C6 in Table 1).

The second scenario (LRS_{25}) assumed a limited dataset of 25 examples instead of 500, with all of them being employed to generate new examples by using LLMs. In this context, 475 examples (C7) were generated, resulting in a total of 500 when combined with the given 25 real sentences, enabling a comparison with the performance archived when using 500 real examples for training (C1). Similarly, the performance was examined when generating an additional 975 synthetic examples (C8) and when generating an additional 1,975 synthetic examples (C9).

3.2. Exemplary low-resource scenario and dataset

To demonstrate the augmentation with LLMs, we considered a low-resource scenario focusing on a German-language domain, specifically German restaurant reviews. As already stated by Hellwig, Fehle, Bink, and Wolff (2024) and Fehle et al. (2023), there are only a few German-language datasets for ABSA to date. We utilized GERestaurant from Hellwig et al. (2024), comprising 3,078 annotated sentences extracted from restaurant reviews. The annotations encompass both implicit and explicit aspects, including their corresponding aspect terms, aspect categories (`GENERAL-IMPRESSION`, `FOOD`, `SERVICE`, `AMBIENCE`, `PRICE`), and sentiment polarity (positive, negative, neutral). GERestaurant exhibits an imbalance in regard to the aspect categories and sentiment polarities.

3.3. Evaluation procedure

A random stratified 6-fold cross-validation was carried out for evaluation. Six iterations were chosen instead of the commonly employed five iterations for cross-validation, since the given annotated dataset employed for evaluation provided a sufficient number of examples to support six folds comprising 500 examples each. The evaluation procedure is further illustrated exemplarily for conditions C3, C4, C7 and C8 in Appendix A.

3.4. Large Language Models (LLMs)

In order to generate the synthetic examples, both an open-source LLM, Llama-3-70B (70 billion parameters) (Touvron et al., 2023), and a commercial model, GPT-3.5-turbo[3] (Brown et al., 2020) were utilized. Technical details on the employed LLMs are explained in more detail in Appendix B.

[3] We used GPT-3.5-turbo with a training cut-off date of 13th June 2023 (GPT-3.5-turbo-0613).

3.5. Prompting strategy

In alignment with the prompting strategy suggested by Zhang et al. (2023) and Møller et al. (2023), the employed prompt included components such as a task name, task definition, output format, and a demonstration section containing annotated examples. An example of a prompt is provided in Appendix C.

3.5.1. Formatting of few-shot examples

Each of the 25 annotated few-shot examples comprised a label and a corresponding sentence, as demonstrated in Table 2. The label was structured as an array containing tuples, with each tuple representing an aspect addressed in the text. Each tuple included the aspect category and sentiment polarity associated with that aspect. In instances where an explicit aspect was present, the aspect terms were enclosed with an XML tag. The aspect category and sentiment polarity of an aspect term were indicated using XML attributes. XML tags provided the capability of nesting, enabling multiple combinations of aspect categories and sentiment polarity to be assigned to a single aspect term.

We also considered the approach of predicting both the text and the positions of aspect terms, rather than directly tagging aspect terms within the text using XML tags. However, we decided against this approach, as LLMs have demonstrated inaccuracies in counting words or characters in text (Davis, 2024; McCoy, Yao, Friedman, Hardy, & Griffiths, 2023).

For LRS_{25}, the selection process of the few-shot examples involved initially choosing five unique random examples from a fold comprising 500 annotated examples for each of the five aspect categories (as illustrated in appendix Fig. A.1(c)) resulting in a total of 25 examples. These 25 examples were used to generate all synthetic examples. Notably, due to the possibility of multiple occurrences or repetitions of an aspect category in a label, a given aspect category could occur more than five times across all 25 few-shot examples.

In the case of LRS_{500}, 25 unique few-shot examples were randomly sampled for each prompt from the available 500 annotated examples. As for the other low-resource scenario, five examples were randomly selected for each aspect category, guaranteeing the presence of at least five few-shot examples with the aspect category appearing at least once in its label.

In both low-resource scenarios, the 25 few-shot examples in the prompt were randomly shuffled for each prompt. The same few-shot examples and prompts were used regardless of the LLM employed for generation.

3.5.2. Label associated with example to be generated

At the end of the prompt, a label was given in the same format as in the few-shot examples, for which a sentence had to be generated by the LLM. We ensured that aspects assigned to each possible combination of aspect category and sentiment polarity as evenly as possible in the generated examples in order to obtain a balanced dataset. The process is described in more detail in Appendix D.

By combining the given label with the LLM's generation, we got a training example that included the sentence and triplets, as illustrated in Table 3. An English version of the illustration is provided in Appendix F. The sentence was obtained by removing the XML tags from the generation. The default assumption was that an aspect, represented by a tuple in the label, was an implicit aspect. In case the generation contained an aspect term corresponding to a combination of aspect

Table 2
Formatting of a few-shot example.

```
Label: [('SERVICE', 'NEGATIVE'), ('FOOD', 'POSITIVE')]
Prediction: Alle vom <aspect-term aspect="SERVICE" polarity="POSITIVE">
Personal</aspect-term> sind sehr unfreundlich, aber es war lecker.
```

English translation: "The staff (Ger.:"Personal") is very friendly (Ger.:"freundlich"), but it was delicious (Ger.:"lecker")."

Table 3
Examples for the translation of a generation into a training example.

Label	`[('SERVICE', 'NEGATIVE'), ('FOOD', 'POSITIVE')]`
Generation	`Alle vom <aspect-term aspect='SERVICE' polarity='POSITIVE'> Personal</aspect-term>sind sehr unfreundlich aber es war lecker.`
Sentence	*"Alle vom Personal sind sehr unfreundlich aber es war lecker."*
Triplets	`[('Personal', 'SERVICE', 'NEGATIVE', start=9, end=17), ('NULL', 'FOOD', 'POSITIVE', start=0, end=0)]`

(a) Explicit and implicit aspects in a generation.

beginequation*7pt]

Label	`[('FOOD', 'POSITIVE'), ('FOOD', 'POSITIVE')]`
Generation	`Es war sehr lecker, auch die <aspect-term aspect='FOOD' polarity='POSITIVE'>Pizza</aspect-term> war lecker.`
Sentence	*"Es war sehr lecker, auch die Pizza war lecker."*
Triplets	`[('Pizza', 'FOOD', 'POSITIVE', start=29, end=34), ('NULL', 'FOOD', 'POSITIVE', start=0, end=0)]`

(b) Duplicate tuple in label.

category and sentiment polarity present in the label, it was considered an explicit aspect.

For a generated example, we analysed the presence of a few characteristics (see Appendix G) in terms of format (e.g. correct XML formatting) and language requirements using basic NLP techniques. If one of the criteria was not met, the prompt was executed again.

3.6. Baseline augmentation method

We employed back-translation, as proposed by Liesting et al. (2021) as a baseline data augmentation method for LRS_{500}. Using the *deep_translator* library,[4] the given 500 examples were first translated into English and then back into German. Each back-translated example was subsequently added to the given dataset, resulting in a total of 1,000 training examples.

3.7. Tasks and models

The nine conditions presented in Table 1, with conditions C4-C9 considered for the two employed LLMs, were evaluated across four common ABSA tasks: ACD, ACSA, E2E-ABSA und TASD. We utilized SOTA transformer-based models for all tasks, as previously employed by Hellwig et al. (2024).

As for SemEval-2014, SemEval-2015, and SemEval-2016, the micro-averaged F1 score was used as the primary evaluation metric. Metrics including the macro-averaged F1 score, precision, recall were calculated as well. The reported value of each metric corresponds to its average across all six iterations.

3.7.1. Aspect Category Detection (ACD) and Aspect Category Sentiment Analysis (ACSA)

Similar to prior studies (Fehle et al., 2023; Hellwig et al., 2024), we treated aspect category detection (ACD) and the identification of both aspect categories and the sentiment polarity expressed towards them (ACSA) as multi-label classification tasks. The base model utilized, *gbert-large*[5] (Chan, Schweter, & Möller, 2020) (337 million parameters), developed by deepset, is based on the BERT architecture pre-trained on large-scale German language corpora (Chan et al., 2020). We fine-tuned this model for both classification tasks. The selection and optimization of hyperparameters is explained in detail in Appendix I.

3.7.2. End-to-End ABSA (E2E-ABSA)

Similar to Li, Bing, Zhang, and Lam (2019), the E2E-ABSA was tackled using a BERT model for token classification. Specifically, we employed the *gbert-large* model for this task. E2E-ABSA involved predicting a tag sequence $y = \{y_1, \ldots, y_T\}$, where each tag corresponds to a token in the sentence. The potential values for y_t include B-$\{POS, NEG, NEU\}$, I-$\{POS, NEG, NEU\}$, or O. These tags denote the beginning (B) and inside (I) of an aspect term, along with negative, neutral, or positive sentiment. In addition to B and I, O was used to denote that a token is not part of an aspect term.

For training, we employed binary cross-entropy loss and used the sigmoid function as the activation function. Following the evaluation methodology of Li et al. (2019), we set the learning rate to 2e−5, batch size to 16, and trained the model for 1,500 steps. During evaluation, true positives included all correctly identified pairs of aspect terms and the sentiment polarity expressed towards them, consistent with the approaches of Zhang et al. (2023) and Li et al. (2019).

3.7.3. Target Aspect Sentiment Detection (TASD)

For the TASD task, the paraphrasing approach by Zhang et al. (2021) was employed, specifically the version adapted for (aspect term, aspect category, sentiment polarity)-triplet annotations in the German language introduced by Hellwig et al. (2024).

t5-base[6] (223 million parameters) was employed as the underlying seq2seq model. In terms of training parameters, batch size was set to 16, number of training epochs to 20 and learning rate to 3e−4, similar to Zhang et al. (2021) and Hellwig et al. (2024). For evaluation, true positives encompassed all correctly identified triplets, meaning that all three sentiment elements (aspect term, aspect category and sentiment polarity) were identified correctly.

[4] deeptranslator: https://pypi.org/project/deep-translator/
[5] German BERT large: https://huggingface.co/deepset/gbert-large
[6] t5-base: https://huggingface.co/google-t5/t5-base

Table 4
Average number of tokens in examples generated by LLMs for each low-resource scenario.

Data Source	LRS	Avg. # Tokens in Sentence	Avg. # Tokens in Examples of each Iteration[a]
Real Examples	–	13.12	–
Llama-3-70B	LRS_{25}	10.83	[10.36, 10.95, 10.08, 10.86, 10.79, 11.96] (*SD* = 0.588)
	LRS_{500}	10.36	[10.39, 10.30, 10.13, 10.08, 10.71, 10.54] (*SD* = 0.221)
GPT-3.5-turbo	LRS_{25}	9.92	[9.24, 9.72, 9.10, 10.32, 10.01, 11.14] (*SD* = 0.686)
	LRS_{500}	9.07	[8.97, 9.11, 8.84, 8.87, 9.40, 9.21] (*SD* = 0.197)

[a] The standard deviation is provided within parentheses for the means when avg. number of tokens are calculated individually across all six iterations.

3.8. Human annotations of synthetic examples

In order to assess the quality of the LLM's annotation, e.g. the ability of an LLM in marking all aspect terms in the text or the reliability of addressing aspects specified in the labels, the annotations generated by the LLM were compared with manual annotations by humans.

A total of 2,400 sentences were annotated, with 600 synthetic examples annotated for each considered LLM and low-resource scenario (2 LLMs × 2 low-resource scenarios × 600 synthetic examples). The choice of selecting 600 synthetic examples arises from randomly selecting 100 examples from the synthetic examples generated for each iteration of the cross-validation process. For illustration, in the case of LRS_{25}, this entails selecting 100 examples from the 1,975 generated examples for a single iteration. As a reminder, for each iteration, no more than 1,975 synthetic examples were generated, aligning with the quantity required for condition C9. The approach allowed for the determination of the average quality of examples generated across all six iterations.

The synthetic examples underwent removal of XML tags, ensuring that the annotators only received the sentence for annotation, without the LLM's annotations of aspect terms in order to prevent a potential bias. The annotation was performed by two expert Annotators: First, annotator *A* annotated all sentences independently, and subsequently, each annotation underwent validation by annotator *B*.

Annotator *B* proposed a label different from that assigned by annotator *A* for 30 out of the 2,400 sentences.

4. Results

In this chapter, we report the properties of the generated examples and the performance that could be achieved with the addition of synthetic examples. Examples generated by the two LLMs for the considered low-resource scenarios can be found in the GitHub repository[7] associated with this work.

4.1. Comparison of real and synthetic data

In order to get a more comprehensive understanding of the linguistic variability in the generated examples, several NLP metrics were computed at the sentence-level.

As illustrated in Table 4, synthetic examples comprised fewer tokens on average than real examples (13.12). However, it can be noticed that examples generated by Llama-3-70B, on average, contained more tokens (10.83 & 10.36) than those generated by GPT-3.5-turbo (9.92 & 9.07). When analysing the mean token count across examples generated for each of the six iterations in the cross-validation setting, examples generated for LRS_{25} exhibited a higher token count when employing Llama-3-70B in comparison to GPT-3.5-turbo. A comparison of the individual iterations is feasible, since, the same few-shot examples were used regardless of the LLM employed except for the cases in which new few-shot examples were used (or the given 25 few-shot examples were shuffled in the case of LRS_{25}) after 25 invalid generations.

[7] https://github.com/NilsHellwig/exploring-absa-llm-augmentation

Table 5
Proportion of examples with the first token identified as a determiner (DET) POS tag.

Data source	LRS	% Examples with the first Token Identified as a Determiner (DET) POS Tag
Real Examples	–	27.3%
Llama-3-70B	LRS_{25}	79.7%
	LRS_{500}	78.6%
GPT-3.5-turbo	LRS_{25}	92.4%
	LRS_{500}	91.9%

On the linguistic level, a notable observation is the predominant use of determiners at the beginning of generations as opposed to realistic examples, particularly with GPT-3.5-turbo, but also with Llama-3-70B (see Table 5).

The number of unique sentences, tokens, and lemmas in the synthetic examples, compared to real examples, is presented in Table 6. In order to examine the impact of the number of examples on the linguistic variability, all metrics for 500, 1,000 and 1,500 examples are presented as the mean over the six iterations. Further details on the preparation of data for the linguistic analysis can be found in Appendix J.

Examining the number of unique sentences generated by Llama-3-70B, there are almost exclusively unique sentences in examples generated for both low-resource scenarios, regardless of the considered number of examples. However, the frequency of unique sentences is slightly lower than that of real examples. In contrast, the number of unique sentences in examples generated by GPT-3.5-turbo is much lower. When considering only 500 examples, there are 308 and 318 unique sentences in the case of LRS_{25} and LRS_{500} respectively. When considering 1,500 examples, slightly more than half of the synthesized sentences are unique in case of both low-resource scenarios (778 unique sentences for LRS_{25} and 784 for LRS_{500}).

The number of unique tokens and lemmas in examples generated by GPT-3.5-turbo is consistently lower compared to examples generated by Llama-3-70B at every sample size and regardless of the examined low-resource scenario. However, for examples generated by Llama-3-70B and GPT-3.5-turbo, the count of unique tokens and lemmas is lower when compared to real examples at every sample size and regardless of the examined low-resource scenario.

Overall, for a given LLM and sample size, the metric values are consistently only slightly larger for examples generated for LRS_{500} compared to those generated for LRS_{25}. The only exception is Llama-3-70B, where, in the case of a sample count of 500 and 1,500, the number of unique sentences is higher in examples generated for LRS_{25} compared to those generated for LRS_{500}.

4.2. Human annotations of synthetic examples

In addition to examining the linguistic variability of synthetic examples, the annotation quality of the utilized LLMs generated for both low-resource scenarios was analysed.

As shown in Table 7, synthetic examples where a conflicting sentiment was expressed towards an aspect, resulting in an assigned polarity label of `CONFLICT`, were infrequent in both examples generated by

Table 6
Linguistic variability of synthetic examples: Sentence-level.

Data Source	LRS	# Examples	# Unique Sentences	# Unique Tokens	# Unique Lemmas
Real Examples	-	500	497	1,918	1,493
		1,000	990	3,061	2,349
		1,500	1,481	3,996	3,038
Llama-3-70B	LRS_{25}	500	480	612	480
		1,000	930	893	700
		1,500	1,367	1,108	860
	LRS_{500}	500	478	666	528
		1,000	930	974	762
		1,500	1,364	1,212	941
GPT-3.5-turbo	LRS_{25}	500	308	296	216
		1,000	553	377	275
		1,500	778	440	319
	LRS_{500}	500	318	294	216
		1,000	560	387	280
		1,500	784	454	327

Table 7
Sentiment polarity label conflicts or unannotated aspects in synthetic examples as detected by human annotation.

LLM	LRS	# Examples with Polarity Conflict	# Examples without annotated Aspects
Llama-3-70B	LRS_{25}	8	10
	LRS_{500}	7	15
GPT-3.5-turbo	LRS_{25}	5	1
	LRS_{500}	4	0

Llama-3-70B and GPT-3.5-turbo. Sentences generated by Llama-3-70B more commonly lacked at least one aspect towards which a sentiment is expressed (10 sentences in the case of LRS_{25}, 15 sentences in the case of LRS_{500}). In the case of GPT-3.5-turbo, this phenomenon was exclusively observed for LRS_{25}, only in 1 out of 600 sentences.

As depicted in Table 8(a), Llama-3-70B exhibited a higher performance in terms of the F1 score regarding annotating aspect terms compared to GPT-3.5-turbo. Higher F1 scores could be observed for Llama-3-70B across both low-resource scenarios. Notably, the precision of Llama-3-70B was lower than that of GPT-3.5-turbo, while recall was higher. In contrast to aspect term annotation, when considering both aspect term and sentiment annotation (see Table 8(b)), the F1 score was higher for GPT-3.5-turbo compared to Llama-3-70B.

The performance in exclusively addressing the aspect categories specified in a label within a generated sentence is presented in Table 8(c). With both micro- and macro-averaged F1 scores above 95 for both LRS_{25} and LRS_{500}, GPT-3.5-turbo achieved a very high performance. When examining sentences generated by Llama-3-70B, the micro-averaged F1 scores (LRS_{25}: 85.90, LRS_{500}: 85.86) were lower compared to GPT-3.5-turbo. Additionally, GPT-3.5-turbo outperformed Llama-3-70B in terms of both micro- and macro-averaged F1 score when considering the addressing of all combinations of aspect category and sentiment polarity specified in the label (see Table 8(d)).

Finally, the identification of triplets was analysed, meaning that a triplet generated through the translation process illustrated in Table 3 was identified by manual annotators as well. The results are presented in Table 8(e). The micro-averaged F1 score was the highest for GPT-3.5-turbo, 55.92 for LRS_{25} and 55.38 for LRS_{500}, surpassing Llama-3-70B, where micro-averaged F1 scores of 46.71 (LRS_{25}) and 47.98 (LRS_{500}) were achieved.

4.3. Task performance

4.3.1. Aspect Category Detection (ACD)

The results of the ACD task and all other tasks are presented in Table 9. For LRS_{25}, the micro- and macro-averaged F1 score consistently remained below that achieved when employing exclusively 500, 1,000, or 2,000 real examples for training, regardless of whether Llama-3-70B or GPT-3.5-turbo was employed for example generation and the quantity of examples that were generated.

In the case of Llama-3-70B and LRS_{25}, the best F1 score (micro: 81.33, macro: 80.37) was achieved when adding 475 synthetic examples for training, the highest number of additional synthetic examples tested here. However, there is a small trend that the micro-averaged F1 score decreased by an increasing number of synthetic training examples. Similarly, in the case of GPT-3.5-turbo and LRS_{25}, the best micro-averaged F1 score was achieved when adding 475 synthetic examples for training (micro: 79.80, macro: 79.13) without further improvement by an increasing number of synthetic training examples. However, regardless of the number of examples generated for LRS_{25}, both micro- and macro-averaged F1 scores were higher for Llama-3-70B than for GPT-3.5-turbo.

Adding synthetic examples (including back-translated examples) to the 500 real examples (LRS_{500}) did not improve both micro- and macro-averaged F1 score, regardless of the LLM considered.

When employing Llama-3-70B for example generation, the addition of an increased number of synthetic examples to the existing 500 real examples in LRS_{500} did not lead to an enhanced micro-averaged F1 score across all aspect categories. In the case of GPT-3.5-turbo and LRS_{500}, the micro-averaged F1 score marginally improved only for the `GENERAL-IMPRESSION` and `FOOD` categories, albeit to a minimal extent.

4.3.2. Aspect Category Sentiment Analysis (ACSA)

In the case of LRS_{25} and the ACSA task, when adding synthetic examples to the given 25 real examples, the achieved micro-averaged F1 score is below that achieved when using exclusively 500, 1,000 or 2,000 real examples, regardless of the number of generated examples and the LLM used for synthesis. In the case of LRS_{25} and Llama-3-70B, the best micro-averaged F1 score (66.07) was achieved when employing when adding 1,975 synthetic examples.

When employing GPT-3.5-turbo for LRS_{25}, slightly higher micro-averaged F1 scores were achieved as compared to Llama-3-70B. In the case of GPT-3.5-turbo, a micro-averaged F1 score of 71.71 was achieved when using 1,975 synthetic examples for training. For LRS_{25} and Llama-3-70B, irrespective of the number of synthetic examples used for training, the macro-averaged F1 score was consistently below the performance achieved when using exclusively 500, 1,000 or 2,000 real examples for training.

Table 8
Comparison of LLM annotations with human annotations.

LLM	LRS	F1 Micro	Precision	Recall
Llama-3-70B	LRS_{25}	**84.12**	89.02	**79.74**
	LRS_{500}	81.52	88.65	75.46
GPT-3.5-turbo	LRS_{25}	75.11	91.44	63.72
	LRS_{500}	72.61	**92.44**	59.78

True Positives: All correctly marked aspect terms using XML tags.

(a) Aspect term.

LLM	LRS	F1 Micro	F1 Macro
Llama-3-70B	LRS_{25}	55.73	39.55
	LRS_{500}	56.05	40.11
GPT-3.5-turbo	LRS_{25}	**67.23**	**50.28**
	LRS_{500}	66.12	49.38

True positives: All correctly marked aspect terms where the sentiment expressed towards them has also been correctly specified in the corresponding XML tag.

(b) Aspect term + sentiment polarity.

LLM	LRS	F1 Micro	F1 Macro
Llama-3-70B	LRS_{25}	85.90	85.26
	LRS_{500}	85.86	85.53
GPT-3.5-turbo	LRS_{25}	95.56	95.42
	LRS_{500}	**95.60**	**95.61**

True Positives: All correctly addressed aspect categories.

(c) Aspect category.

LLM	LRS	F1 Micro	F1 Macro
Llama-3-70B	LRS_{25}	58.98	49.72
	LRS_{500}	61.68	48.67
GPT-3.5-turbo	LRS_{25}	**84.37**	71.46
	LRS_{500}	83.66	**77.60**

True Positives: All correctly addressed combinations of aspect categories and sentiment.

(d) Aspect category + sentiment polarity.

LLM	LRS	F1 Micro	F1 Macro
Llama-3-70B	LRS_{25}	46.71	43.94
	LRS_{500}	47.98	45.02
GPT-3.5-turbo	LRS_{25}	**55.92**	**55.30**
	LRS_{500}	55.38	54.71

True Positives: All addressed/marked implicit and explicit aspects (aspect term) where its assigned aspect category and sentiment polarity are given in the label.

(e) Triplets: Aspect term + aspect category + sentiment polarity.

When generating 975 examples using GPT-3.5-turbo, a macro-averaged F1 score of 62.56 was achieved, and a macro-averaged F1 score of 63.67 in the case of 1,975 synthetic training examples. In contrast to the score obtained when adding 475 synthetic examples to training (50.86), both scores seem higher than that achieved when using exclusively 500 real training examples (59.52).

Finally, in contrast to the ACD task, when employing Llama-3-70B and GPT-3.5-turbo for LRS_{25}, the micro- and macro-averaged F1 score improved with an increasing number of synthetic examples used for training.

In the context of LRS_{500}, the micro-averaged F1 score was not improved by adding synthetic examples generated by Llama-3-70B. However, when adding 500, 1,000, or 1,500 synthetic examples generated by Llama-3-70B in addition to the given 500 real examples for training, the macro-averaged F1 scores were consistently greater than that observed when using exclusively 500 real examples for training. The macro-averaged F1 scores are by far above that, achieved when adding the back-translated examples.

In the case of GPT-3.5-turbo and LRS_{500}, the micro-averaged F1 score exhibited slight improvements by adding 500 (86.70), 1,000 (86.60), or 1,500 (85.94) synthetic examples to the given 500 real examples for training. The inclusion of back-translated examples improved performance to a much smaller extent. A notable improvement in the macro-averaged F1 score was observed, particularly in the case of GTP-3.5-turbo and LRS_{500}. By adding 500, 1,000, or 1,500 examples generated by GPT-3.5-turbo for the training, the macro-averaged F1 score was much higher than that achieved with exclusively 500 real training examples.

In a next step, we observed that the macro-averaged F1 score is higher with 500, 1,000, or 1,500 synthetic examples being included in the training set along the given 500, compared to using exclusively 1,000 real examples for training. Wenn adding 500 synthetic examples to the given 500 real examples, the F1 macro is even above that achieved when using exclusively 2,000 real examples for training. For GPT-3.5-turbo and LRS_{500}, the micro- and macro-averaged F1 score did not increase with an increasing number of synthetic examples.

The F1 scores for all prediction classes (15 in total: 5 aspect categories × 3 sentiment polarities) play a role in elucidating the observed enhancement in the macro-averaged F1 scores when incorporating synthetic examples in the aforementioned scenarios. Notably, when

Table 9
Performance scores for ACD, ACSA, E2E and TASD: Micro- and macro-averaged F1 scores for ABSA models trained for LRS_{25} and LRS_{500}.

Data Source	LRS	# Real	# Synth	ACD		ACSA		E2E		TASD	
				$F1_{Micro}$	$F1_{Macro}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{Micro}$	$F1_{Macro}$
Real Examples	-	500	0	$90.90_{1.37}$	$89.97_{1.66}$	$84.54_{1.14}$	$59.52_{1.66}$	$77.16_{1.77}$	$70.07_{5.01}$	$61.80_{1.77}$	$53.03_{4.86}$
		1,000	0	$92.02_{1.19}$	$91.10_{1.46}$	$88.60_{1.29}$	$74.64_{4.26}$	$80.69_{1.65}$	$75.03_{2.61}$	$65.44_{1.05}$	$58.29_{6.31}$
		2,000	0	$\mathbf{92.35}_{1.15}$	$\mathbf{91.53}_{1.31}$	$\mathbf{89.40}_{1.37}$	$\mathbf{78.86}_{5.06}$	$\mathbf{82.00}_{3.68}$	$\mathbf{78.83}_{4.89}$	$\mathbf{68.96}_{1.31}$	$\mathbf{60.22}_{4.53}$
Back-translation	LRS_{500}	500	500	$90.49_{0.78}$	$89.25_{1.11}$	$85.32_{1.28}$	$64.42_{3.34}$	$78.08_{2.08}$	$70.41_{6.21}$	$62.92_{1.58}$	$54.93_{4.85}$
Llama-3-70B	LRS_{25}	25	475	$\mathbf{81.33}_{0.36}$	$\mathbf{80.37}_{1.30}$	$60.68_{4.27}$	$49.93_{3.41}$	$51.95_{2.41}$	$44.50_{2.00}$	$38.53_{3.89}$	$29.84_{2.36}$
		25	975	$80.76_{2.03}$	$80.18_{1.65}$	$64.95_{3.41}$	$54.42_{3.60}$	$53.65_{5.49}$	$46.81_{5.40}$	$39.10_{3.01}$	$\mathbf{30.27}_{2.02}$
		25	1,975	$80.65_{1.77}$	$80.20_{1.11}$	$\mathbf{66.07}_{3.95}$	$\mathbf{55.58}_{3.52}$	$\mathbf{57.34}_{3.82}$	$\mathbf{48.52}_{3.10}$	$\mathbf{39.14}_{2.35}$	$29.50_{1.84}$
	LRS_{500}	500	500	$\mathbf{89.88}_{1.42}$	$\mathbf{88.77}_{1.72}$	$\mathbf{83.22}_{1.00}$	$\mathbf{70.66}_{3.81}$	$\mathbf{75.25}_{1.59}$	$\mathbf{66.86}_{2.23}$	$58.33_{1.52}$	$44.32_{3.23}$
		500	1,000	$88.77_{1.06}$	$87.15_{1.10}$	$82.29_{1.30}$	$67.96_{4.30}$	$73.19_{2.89}$	$63.80_{2.49}$	$\mathbf{58.66}_{1.10}$	$\mathbf{44.66}_{3.04}$
		500	1,500	$88.49_{1.61}$	$87.11_{1.55}$	$80.64_{1.50}$	$66.52_{3.84}$	$71.98_{3.01}$	$61.95_{2.57}$	$56.23_{1.98}$	$41.49_{2.06}$
GPT-3.5-turbo	LRS_{25}	25	475	$\mathbf{79.80}_{2.57}$	$\mathbf{79.13}_{2.15}$	$60.18_{6.20}$	$50.86_{5.30}$	$59.15_{4.39}$	$52.61_{5.08}$	$36.96_{2.60}$	$29.25_{4.49}$
		25	975	$79.52_{3.32}$	$78.80_{3.58}$	$70.95_{4.35}$	$62.56_{5.60}$	$58.65_{4.70}$	$52.66_{5.71}$	$\mathbf{37.88}_{3.95}$	$\mathbf{31.53}_{5.20}$
		25	1,975	$79.63_{3.17}$	$79.11_{3.49}$	$\mathbf{71.71}_{3.59}$	$\mathbf{63.67}_{3.05}$	$\mathbf{61.32}_{3.87}$	$\mathbf{58.15}_{3.81}$	$37.39_{3.13}$	$28.74_{3.55}$
	LRS_{500}	500	500	$\mathbf{90.85}_{0.91}$	$\mathbf{89.89}_{1.16}$	$\mathbf{86.70}_{2.03}$	$\mathbf{79.00}_{7.11}$	$\mathbf{76.79}_{3.89}$	$72.27_{3.78}$	$\mathbf{61.39}_{1.15}$	$\mathbf{50.83}_{6.11}$
		500	1,000	$90.52_{1.33}$	$89.81_{1.42}$	$86.60_{1.36}$	$78.42_{5.28}$	$75.96_{3.26}$	$\mathbf{72.55}_{2.91}$	$59.51_{1.90}$	$50.51_{6.63}$
		500	1,500	$89.68_{0.96}$	$88.77_{1.23}$	$85.94_{1.45}$	$78.47_{3.31}$	$76.14_{3.54}$	$71.67_{3.02}$	$59.42_{1.74}$	$48.76_{5.04}$

Mean was calculated for all metrics, derived from six iterations of the cross-validation and are accompanied by the standard deviation for each metric, calculated across the six values.

relying exclusively on 500 real examples for training, the F1 score was 0 for the recognition of a neutral sentiment in the aspect categories `GENERAL-IMPRESSION`, `SERVICE`, `AMBIENCE`, and `PRICE`, denoting an absence of true positives. The augmentation with synthetic examples notably boosted the F1 score for these four classes, associated with a neutral sentiment, evident for both LRS_{25} and LRS_{500}.

4.3.3. End-to-End Aspect Based Sentiment Analysis (E2E-ABSA)

In the context of LRS_{25}, the micro- and macro-averaged F1 score using examples generated by Llama-3-70B and GPT-3.5-turbo was lower than that achieved using exclusively 500, 1,000 or 2,000 real training examples. Furthermore, an improved micro- and macro-averaged F1 score was achieved with an increasing number of synthetic examples. For LRS_{25}, the best F1 score (micro: 61.32, macro: 58.15) was observed when adding 1,975 synthetic examples generated by GPT-3.5-turbo to the given 25 real examples.

Neither adding synthetic sentences generated by GPT-3.5-turbo nor those generated by Llama-3-70B led to an improvement in the micro-averaged F1 score for LRS_{500} compared to using only 500 real examples for training. Furthermore, in the case of LRS_{500}, no improvement of the micro-averaged F1 score was observed with an increased number of synthetic examples. Conversely, adding 500 back-translated examples to the given 500 real examples led to an improvement of the F1 micro to 78.08.

An improvement in the macro-averaged F1 score could not be achieved through the inclusion of examples generated by Llama-3-70B in the case of LRS_{500}. When augmenting the existing 500 examples with those generated by GPT-3.5-turbo for training, the macro-averaged F1 score showed enhancement, even higher than that achieved when adding back-translated examples, reaching 72.27 for 500, 72.55 for 1,000, and 71.67 for 1,500 synthetic examples.

Looking at the performance achieved with respect to each class, we noticed that in the case of LRS_{500}, an improvement of the F1 score was achieved for the neutral polarity when adding examples generated with GPT-3.5-turbo.

4.3.4. Target Aspect Sentiment Detection (TASD)

In the case of LRS_{25}, the addition of 475, 975, and 1,975 synthetic examples to the existing 25 real examples did not yield a micro-averaged F1 score superior to that achieved with exclusively 500, 1,000, or 2,000 real examples, but even resulted in a much lower score. For LRS_{25} and Llama-3-70B, the highest micro-averaged F1 score, 39.14, was reported when incorporating 1,975 synthetic examples for training. In case GPT-3.5-turbo-generated examples were introduced for LRS_{25}, a micro-averaged F1 score of up to 37.88 could be achieved, specifically with the addition of 975 synthetic examples.

Concerning LRS_{500}, there was no improvement in the F1 score when incorporating synthetic examples, regardless of the number of synthetic training examples considered and the specific LLM used for example generation. Overall, for both low-resource scenarios, no clear trend that increasing the number of synthetic examples lead to an increase in the F1 score was observed in either the case of Llama-3-70B or GPT-3.5-turbo. Applying Back-translation, however, allowed for a boost in both the micro- and macro-averaged F1 score.

Regardless of the LLM used for generating annotated examples, the low-resource scenario considered, and the number of synthetic examples added to the training set, the macro-averaged F1 score never exceeded that achieved using exclusively 500, 1,000, or 2,000 real examples.

5. Discussion

This work employed LLMs for generating annotated examples for ABSA. An LLM was used to generate sentences with annotations of aspect terms, their corresponding aspect category and the sentiment expressed towards them. Low-resource scenarios assuming the availability of only 25 (LRS_{25}) or 500 real examples (LRS_{500}) were considered. In the case of LRS_{25}, each synthetic example was generated using all available 25 examples as few-shot examples in the prompt, whereas in the case of LRS_{500}, 25 out of 500 given examples were randomly selected as few-shot examples for the generation of each synthetic example.

5.1. Generation of annotated examples with Llama-3-70B and GPT-3.5-turbo

For the examined low-resource scenarios and employed LLMs, there was often a need to re-execute the prompt due to the generated example not exhibiting the desired characteristics. In order to reduce the number of regenerations and therefore computation time or monetary costs, one approach could be to provide more detailed guidance on these characteristics in the task description of the prompt. In the case of LRS_{500}, using more than 25 examples from the available set of 500 real examples as few-shot examples in the LLM's prompt could be explored in order to improve the LLMs understanding of the required format of an annotated example.

5.2. Disparities in linguistic variability and annotation quality across examined LLMs

For both LRS_{25} and LRS_{500}, examples generated by Llama-3-70B demonstrated a higher linguistic variability than those generated by GPT-3.5-turbo in terms of metrics such as the number of unique sentences, tokens, and lemmas. Since we did not evaluate multiple hyperparameter configurations of the LLMs for the generation of our synthetic examples, the linguistic variability in synthetic sentences by GPT-3.5-turbo may be enhanced by a different specification of hyperparameters of the LLM's such as its temperature.

Furthermore, 2,400 synthetic examples were annotated by human annotators. Llama-3-70B outperforms GPT-3.5-turbo in annotating all aspect terms in the examples. However, if the aim is to mark both all aspect terms and assigning the sentiment expressed towards them, the micro-averaged F1 score is higher for GPT-3.5-turbo. GPT-3.5-turbo also exhibits a higher micro-averaged F1 score in addressing exactly the aspect categories (and sentiment polarities) specified in the label. Finally, the identification of triplets was scrutinized by examining whether manually annotated triplets aligned with those generated through the translation process depicted in Table 3. In this context, GPT-3.5-turbo demonstrated a higher micro- and macro-averaged F1 score compared to Llama-3-70B.

Considering the differing annotation quality and language variability across the LLMs employed, one could explore further LLMs for example generation. This exploration may encompass LLMs with different parameter sizes, context sizes and LLMs explicitly designed for the language under consideration (in our application, German). Moreover, considering LRS_{500}, fine-tuning an LLM on the available 500 real examples could be pursued, enabling the LLM to capture insights from all 500 available annotated real examples.

Finally, given the higher linguistic variability observed in examples generated by Llama-3-70B compared to those generated by GPT-3.5-turbo as well as the superior performance of GPT-3.5-turbo in annotating examples, it could be considered to employ two distinct LLMs for the processes of text generation and annotation, respectively. Notably, this approach would result in increased computational or financial costs, as each step would necessitate the execution of a distinct LLM for generating each example.

5.3. Augmentation with synthetic examples in low-resource scenarios

The generated examples were finally utilized to train SOTA SLMs based on the transformer-architecture dedicated to four common ABSA tasks. Notably, SLMs demand less computational power than LLMs, allowing them to proficiently handle individual ABSA tasks with reduced computational requirements.

5.3.1. Performance in LRS_{25}

Looking at the overall task performances, irrespective of the LLM and the quantity of synthetic examples considered for LRS_{25}, a superior micro-averaged F1 score was never achieved compared to using exclusively 500, 1,000, or 1,500 real training examples. The micro-averaged F1 score reached its peak in all but three cases when 1,975 synthetic training examples were added to the given 25 real examples. For the ACD task, employing both Llama-3-70B and GPT-3.5-turbo for example generation yielded best micro-averaged F1 scores when utilizing only 475 generated examples for training, in addition to the existing 25 real examples. Furthermore, in the case of GPT-3.5-turbo and the TASD task, the best micro-averaged F1 score was achieved when adding 975 examples. However, in all three cases, the F1 micro score was only slightly lower when adding more synthetic training examples.

A limitation of this study may be that no investigation of a larger set of synthetic examples for training was conducted beyond 1,975 synthetic examples. This potentially allows performance enhancements and an investigation of the threshold at which adding more synthetic examples no longer contributes to performance improvement.

When evaluating the ACD task, micro-averaged F1 scores of 81.33 and 79.80 were achieved for examples generated by Llama-3-70B and GPT-3.5-turbo, respectively. In the ACSA task, a notable performance gap between the evaluated LLMs was observed, with Llama-3-70B achieving a micro-averaged F1 score of 66.07, which was lower than the micro-averaged F1 score of 71.71 achieved when adding training examples generated by GPT-3.5-turbo.

The E2E-ABSA task yielded micro-averaged F1 scores of 57.34 for Llama-3-70B and 61.32 for GPT-3.5-turbo. This performance aligns closely with results reported by Zhang et al. (2023), who achieved comparable scores using ChatGPT as a classifier for the E2E-ABSA task with a prompt comprising 10 few-shot examples per sentiment polarity (Rest14: 63.3, Rest15: 52.85, Rest16: 59.22).

Finally, in the TASD task, a substantial difference in results emerged between using exclusively real training examples and synthetic examples generated for LRS_{25}. Employing Llama-3-70B-generated training examples yielded a micro-averaged F1 score of 39.14, whereas the use of GPT-3.5-turbo-generated examples resulted in a score of 37.88. Conversely, utilizing exclusively 500 real examples led to a higher performance, a micro-averaged F1 score of 61.80.

5.3.2. Performance in LRS_{500}

In the case of LRS_{500}, the inclusion of LLM-generated examples alongside the existing 500 real examples for training did not improve the micro-averaged F1 score, while adding back-translated examples did boost performance for all tasks except for the ACD task. An exception was the ACSA task, where the micro-averaged F1 score exhibited improvement with the incorporation of 500, 1,000 or 1,500 examples generated by GPT-3.5-turbo. While exclusively using real examples resulted in an average F1 micro score of 84.54, introducing 500 synthetic examples improved the micro-averaged F1 score to 86.70. Notably, the boost in performance was higher than that observed when adding 500 back-translated examples (85.32).

In the same way, the macro-averaged F1 score demonstrated an improvement as well, when 500, 1,000 or 1,500 examples generated by Llama-3-70B or GPT-3.5-turbo were added to the existing 500 real training examples. It increased from 59.52 when using exclusively 500 real training examples to 79.00 when 500 examples generated by GPT-3.5-turbo were incorporated alongside the 500 real examples.

This improvement can be attributed to the equal frequency of each sentiment polarity in 500, 1,000 or 1,500 synthetic examples. This stands in contrast to the real examples, where a neutral sentiment is infrequent compared to positive and negative sentiments. In four of the examined classes in the ACSA task (15 in total: 5 aspect categories × 3 sentiment polarities) associated with a neutral sentiment, reported F1 scores were 0. Through the inclusion of synthetic examples, these scores could be enhanced. Notably, the performance boost of the F1 macro score is far beyond that, observed when adding back-translated examples (64.42), underlining the huge potential of LLMs for improving the performance for rare classes.

In the case of the E2E-ABSA task, an enhancement in the macro-averaged F1 score was observed by adding 500, 1,000, or 1,500 examples. While achieving a score of 70.07 using solely 500 real training examples, the highest improvement was attained by adding 1,000 GPT-3.5-turbo-generated examples (72.55).

One limitation of the study might be the lack of an evaluation of a balanced training approach, wherein, for example, sentences with a neutral sentiment receive additional weight during the optimization of model parameters due to their scarcity among real data. It could be investigated whether the inclusion of synthetic examples still leads to an improvement when being compared to a balanced-training approach. Finally, in the context of LRS_{500}, an even higher improvement in performance might be achieved by applying the aforementioned measures to improve linguistic variability and annotation quality.

6. Conclusion and future work

In the present work, two LLMs, namely Llama-3-70B and GPT-3.5-turbo, were utilized to generate annotated training examples for ABSA, encompassing annotations of aspect terms, aspect categories, and sentiment polarities.

The study explored two low-resource scenarios, LRS_{25} and LRS_{500}, considering the availability of a pool of 25 or 500 annotated real examples, respectively. LLMs were instructed with prompts containing a task description and 25 few-shot examples randomly drawn from the pool to generate additional annotated sentences. At the end of the prompt, a label was provided, consisting of one or more tuples, each representing an aspect. A tuple comprised an aspect category of an aspect to be discussed in the generated sentence, and a sentiment polarity to be expressed towards the aspect.

The results revealed that the generated examples showed lower linguistic variability in terms of unique sentences, tokens, and aspect terms compared to real examples. In comparison, Llama-3-70B-generated examples demonstrated much greater variability than those generated by GPT-3.5-turbo. In the next step, transformer-based models with less than 1 billion parameters were trained on synthetic examples added to the given real examples. In the case of LRS_{25}, a high micro-averaged F1 score (81.33) was achieved in the ACD task, while only a micro-averaged F1 score of 39.14 was achieved in the TASD task. In the case of LRS_{500} and ACSA, synthetic examples generated with GPT-3.5-turbo improved both the micro- and macro-averaged F1 scores.

Future work could explore generating annotated examples for domains and languages other than that considered in this study. Moreover, given the observed difference in performance depending on the LLM employed for augmentation, future work could explore other LLMs with varying parameter sizes and training data used for their pre-training. This may include LLMs pre-trained on texts specific to the domain or language of interest. In order to address time and financial constraints, exploring methods to generate multiple annotated examples with a single LLM execution as performed by Møller et al. (2023) could be explored.

Akin to the SemEval datasets (Pontiki et al., 2016, 2015, 2014), training examples were generated along with annotations for aspect terms, aspect categories, and sentiment polarity. One could explore generating training examples that only include annotations for the sentiment elements required for the corresponding ABSA task. Subsequently, it could be assessed whether utilizing these examples further improves performance in the corresponding task. Furthermore, the capability of LLMs in annotating opinion terms in generated examples could be investigated, a sentiment element that was not considered in this study.

Finally, future work could investigate the task performance achievable when annotated examples are generated without the presence of few-shot examples (zero-shot learning).

CRediT authorship contribution statement

Nils Constantin Hellwig: Writing – original draft, Methodology, Data curation, Software. **Jakob Fehle:** Writing – original draft, Methodology, Data curation, Project administration, Software. **Christian Wolff:** Supervision, Writing – review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Training and evaluation workflow

See Fig. A.1.

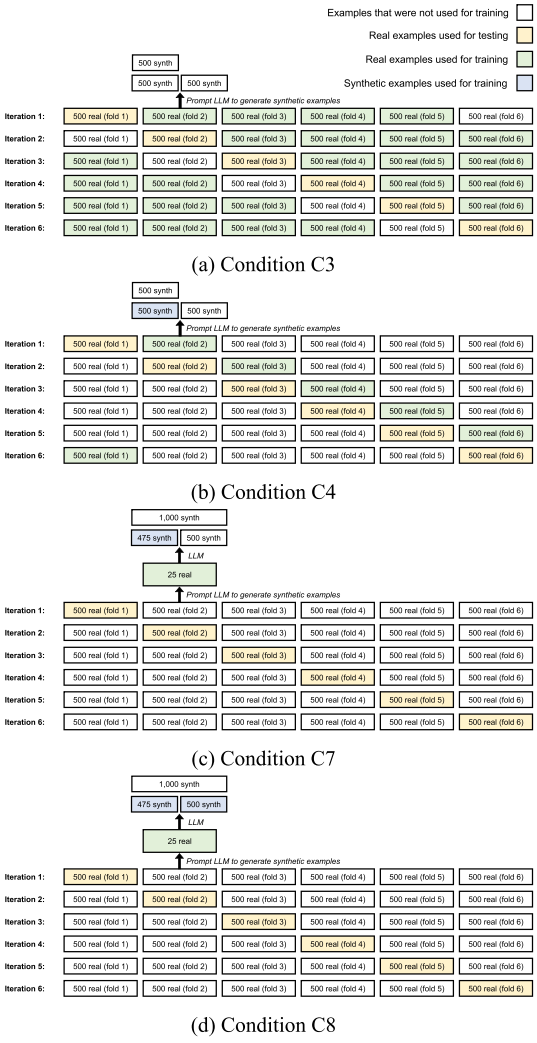


Fig. A.1. Dataset-splitting to perform a six-fold cross-validation: Illustration of employed training samples for each iteration using condition C3, C4, C7 and C8 as an example.

Appendix B. Technical details on the employed LLMs

B.1. Llama-3-70B

Due to its huge size, Llama-3-70B[8] could not be loaded onto the NVIDIA Quadro RTX 6000 GPU (24 GB GDDR6 GPU memory) used for this study. Consequently, a 4-bit quantized version of the LLM[9] was utilized. The required memory is lowered by reducing the number of bits required to represent each parameter of the LLM (Dettmers & Zettlemoyer, 2023). For inference of the quantized model, *llama.cpp*[10] was used, specifically its official Python bindings.[11]

B.2. GPT-3.5-turbo

In order to leverage GPT-3.5-turbo,[12] the OpenAI Python package[13] was utilized. All requests to the OpenAI API and inference of

[8] meta-llama/Meta-Llama-3-70B: https://huggingface.co/meta-llama/Meta-Llama-3-70B

[9] NousResearch/Meta-Llama-3-70B-GGUF: https://huggingface.co/NousResearch/Meta-Llama-3-70B-GGUF

[10] *llama.cpp*: https://github.com/ggerganov/llama.cpp

[11] llama-cpp-python: https://pypi.org/project/llama-cpp-python

[12] OpenAI - GPT-3.5: https://platform.openai.com/docs/models/gpt-3-5

```
Erzeuge genau einen Satz einer deutschsprachigen Restaurant-Bewertung, der für das Training eines Modells
für die Aspekt-basierte Sentiment Analyse verwendet werden kann.
Gegeben ist ein Label in Form eines Arrays, wobei ein oder mehrere Tupel (Aspekt-Kategorie, Sentiment-
Polarität) gegeben sind.
Für ein Label wird eine Prediction erzeugt, wobei die Prediction ausschließlich die in dem Label
definierten Kombinationen aus Aspekt-Kategorie und Aspekt-Polarität adressiert.

Folgende Aspekt-Kategorien werden betrachtet:

* "FOOD" - Aspekte in Bezug auf das Essen im allgemeinen oder bestimmte Speisen und Getränke
* "SERVICE" - Aspekte in Bezug auf den Service im allgemeinen oder Einstellung und Professionalität des
Personals, die Wartezeiten oder Service-Dienstleistungen wie Speisenmitnahme
* "PRICE" - Aspekte in Bezug auf den Preis im allgemeinen oder Speisen, Getränke oder andere Leistungen des
Restaurants, deren Preis bewertet wird.
* "AMBIENCE" - Aspekte in Bezug auf das Ambiente und Atmosphäre im allgemeinen oder die Umgebung des Innen-
und Außenbereichs des Restaurants, dessen Ausstattung und Unterhaltungsmöglichkeiten
* "GENERAL-IMPRESSION" - Aspekte in Bezug auf das Restaurant als Ganzes ohne Fokus auf die oben genannten
Aspekt-Kategorien

Gegenüber einem Aspekt können folgende Stimmungen geäußert werden:

* "POSITIVE"
* "NEUTRAL"
* "NEGATIVE"

Zusätzlich kann für ein im Label vorgegebenes Tupel (Aspekt-Kategorie, Aspekt-Polarität) ein Aspekt-Begriff
im Text vorliegen.
Ein Aspekt-Begriff ist eine Entität oder Eigenschaft innerhalb eines Textes, die auf eine der betrachteten
Aspekt-Kategorien hinweist.
Indikatoren für Subjektivität (d.h. Wörter oder Phrasen, die Meinungen, Bewertungen usw. ausdrücken) gelten
nicht als Aspekt-Begriffe oder Bestandteile von Aspekt-Begriffen.
Wichtig: Aspekt-Begriffe werden immer im Text mit einem XML-Tag markiert!
Ein im Tupel vorgegebener Aspekt kann im Text auch implizit adressiert werden, wobei für einen solchen
Aspekt kein Aspekt-Begriff im Text mit einem XML-Tag markiert wird.

Gebe nur die Prediction zurück, ohne Kommentare und ohne zusätzlichen Text.

Label:[('FOOD', 'NEGATIVE')]
Prediction:Leider hat das <aspect-term aspect="FOOD" polarity="NEGATIVE">Essen</aspect-term> nicht unseren
Erwartungen entsprochen.
[[... 24 examples]]
Label:[('GENERAL-IMPRESSION', 'POSITIVE')]
Prediction:
```

Fig. C.2. Prompt for generating an annotated example: Each prompt comprises a task description, annotated examples and a label. An example appropriate to the label should be generated.

Llama-3-70B were executed sequentially without any parallelization. The maximum context length was 4,096 and 8,192 tokens, for GPT-3.5-turbo and Llama-3-70B respectively, which was sufficient for all prompts used in this study. The temperature parameter was set to 0.5 for both LLMs in order to balance variability and coherence. The termination criterion for token generation was specified as "\n", representing a newline character, since synthetic examples, similar to the few-shot examples, were required to be consistently presented in a single line.

Appendix C. Examples of prompts

C.0.1. Example

See Fig. C.2.

C.0.2. Example translated to English

See Fig. C.3.

Appendix D. Specification of labels of the examples to be generated

The label followed the identical format as utilized for the few-shot examples. The generation of examples for LRS_{500} was executed in such a way that for synthetic examples, the frequency of a given number of tuples in the label corresponds to the frequency in the available 500 real examples, which served as a representation of the overall population. For example, if 80% of the generated examples contained only one tuple in their corresponding label, it would be because 400 out of the 500 real examples also had only one tuple in their label (e.g. 400 of 500 examples if 500 synthetic examples are to be generated).

For LRS_{25}, this modelling was not possible, since the quantity of 25 annotated real examples is too small to make a representative judgment about the overall population. Notably, in this low-resource scenario, there is no access to the 500 real examples, preventing their distribution from serving as a reference.

Consequently, for such a low-resource scenario, where no representative data is available, a distinct strategy was employed. All training sets from the ABSA task in SemEval-2016, encompassing customer reviews in various languages and from different domains, were utilized to determine the distribution used for the labels. Thus, the distribution was derived from the distribution observed in these examples. Notably, the distribution of the number of tuples in a label, which was used to generate examples, is shown in Appendix E for both low-resource scenarios. Interestingly, the distribution calculated based on the SemEval datasets was very similar to the distributions calculated for the six iterations based on the given 500 examples available in each iteration.

It was ensured, that each combination of aspect category and sentiment polarity (15 in total since there are 5 aspect categories and 3 sentiment polarities) occurred with equal frequency among the labels for the examples to be generated. For instance, the tuple `('FOOD','POSITIVE')` had to occur with the same frequency as the tuple `('SERVICE','NEUTRAL')`. As the number of tuples across all labels was not always divisible by 15, exact equality was sometimes impossible. This approach for label balancing allowed for the inclusion of combinations of aspect categories and sentiment polarities that may only rarely have occurred in the real examples.

Appendix E. Distribution of the number of tuples in a label for few-shot generation

See Table E.10.

[13] openai: https://pypi.org/project/openai

```
Generate exactly one sentence of a German restaurant review that can be used for training a model for
aspect-based sentiment analysis. Given is a label in the form of an array, where one or more tuples (aspect
category, sentiment polarity) are provided. For a label, a prediction is generated, whereby the prediction
addresses exclusively the combinations of aspect category and aspect polarity defined in the label.

The following aspect categories are considered:

* "FOOD" - Aspects related to food in general or specific dishes and drinks
* "SERVICE" - Aspects related to service in general or attitude and professionalism of the staff, waiting
times, or service offerings like take-out
* "PRICE" - Aspects related to price in general or restaurant's dishes, drinks, or other services whose
price is evaluated.
* "AMBIENCE" - Aspects related to the ambiance and atmosphere in general or the environment of the interior
and exterior of the restaurant, its decor, and entertainment options
* "GENERAL-IMPRESSION" - Aspects related to the restaurant as a whole without focusing on the above-
mentioned aspect categories

The following sentiments can be expressed towards an aspect:

* "POSITIVE"
* "NEUTRAL"
* "NEGATIVE"

Additionally, for a given tuple (aspect category, aspect polarity) in the label, an aspect term may be
present in the text. An aspect term is an entity or property within a text that refers to one of the
considered aspect categories. Indicators of subjectivity (i.e., words or phrases expressing opinions,
evaluations, etc.) are not considered aspect terms or components of aspect terms. Important: Aspect terms
are always marked in the text with an XML tag! An aspect specified in the tuple can also be implicitly
addressed in the text, with no aspect term marked in the text with an XML tag for such an aspect.

Return only the prediction, without comments and without additional text.

Label:[('FOOD', 'NEGATIVE')]
Prediction:Unfortunately, the <aspect-term aspect="FOOD" polarity="NEGATIVE">food</aspect-term> didn't meet
our expectations.
[[... 24 examples]]
Label:[('GENERAL-IMPRESSION', 'POSITIVE')]
Prediction:
```

Fig. C.3. Prompt for generating an annotated example (English translation).

Table E.10
Distribution of the Number of Tuples in a Label for Few-Shot Generation.

# Tuples in Label	1	2	3	4	5	6	7	8	9	16
SemEval Ratio for LRS_{25}	64.8%	21.0%	8.2%	3.6%	1.4%	0.6%	0.2%	0.2%	–	–
Iteration 1 for LRS_{500}	72.6%	18.8%	5.2%	2.2%	0.4%	0.6%	0.2%	–	–	–
Iteration 2 for LRS_{500}	73.8%	18.2%	5.0%	2.4%	0.2%	0.2%	–	–	0.2%	–
Iteration 3 for LRS_{500}	73.6%	18.6%	5.4%	1.0%	0.8%	0.2%	0.4%	–	–	–
Iteration 4 for LRS_{500}	73.8%	20.0%	4.4%	1.6%	0.2%	–	–	–	–	–
Iteration 5 for LRS_{500}	69.6%	20.0%	6.4%	2.6%	0.6%	0.4%	–	0.2%	–	0.2%
Iteration 6 for LRS_{500}	72.0%	19.6%	6.4%	1.4%	0.6%	–	–	–	–	–

Table F.11
Examples for the translation of a generation into a training example.

Label	[('SERVICE', 'NEGATIVE'), ('FOOD', 'POSITIVE')]
Generation	The <aspect-term aspect='SERVICE' polarity='POSITIVE'>staff </aspect-term> is very unfriendly, but it was delicious.
Sentence	*"The staff is very unfriendly, but it was delicious."*
Triplets	[('staff', 'SERVICE', 'NEGATIVE', start=4, end=9), ('NULL', 'FOOD', 'POSITIVE', start=0, end=0)]

(a) Explicit and implicit aspects in a generation

Label	[('FOOD', 'POSITIVE'), ('FOOD', 'POSITIVE')]
Generation	It was very delicious, even the <aspect-term aspect='FOOD' polarity='POSITIVE'>pizza</aspect-term> was delicious.
Sentence	*"It was very delicious, even the pizza was delicious."*
Triplets	('pizza', 'FOOD', 'POSITIVE', start=32, end=37), ('NULL', 'FOOD', 'POSITIVE', start=0, end=0)]

(b) Duplicate tuple in label

Appendix F. Examples for the translation of a generation into a training example: English example

See Table F.11.

Appendix G. Characteristics analysed in the validation process

See Table G.12.

Characteristics that must be present in a generated example:

- XML tags for marking aspect terms must have a valid XML scheme (e.g., closing XML tag is present).
- Exclusivity of XML tags named "`aspect-term`", each featuring attributes "`aspect`" and "`polarity`". These attributes are strictly defined to encompass only the five introduced aspect categories and three sentiment polarities as their respective values.

Table G.12
Validation process of a marked aspect term.

Generations for Label `[('FOOD', 'POSITIVE')]`
"Die <aspect-term aspect='FOOD' polarity='POSITIVE'>Pizza</aspect-term> war gut."
✓ Valid Generation: `('FOOD', 'POSITIVE')` marking the aspect term `'Pizza'` occurs in the given label.
"Es hat mir gut geschmeckt!"
✓ Valid Generation: There is no aspect term marked in the generated text that has been assigned a combination of aspect category and sentiment polarity not present in the label.
"Die <aspect-term aspect='FOOD' polarity='POSITIVE'>Pizza</aspect-term> hat mir nicht geschmeckt."
✓ Valid Generation: `'Pizza'` is assigned the aspect category FOOD and the sentiment polarity `'POSITIVE'` which is also specified in the label. Negative sentiment expressed towards `'Pizza'` is not verified automatically, since the validation process operates at a syntactical level rather than a semantic!
"Die <aspect-term aspect='FOOD' polarity='POSITIVE'>Pizza</aspect-term> und das <aspect-term aspect='FOOD' polarity='POSITIVE'>Eis</aspect-term>haben mir gut geschmeckt."
✗ Invalid Generation: `('FOOD', 'POSITIVE')` occurs only once in the provided label.
"Die <aspect-term aspect='FOOD' polarity='NEGATIVE'>Pizza</aspect-term> war gut."
✗ Invalid Generation: `('FOOD', 'NEGATIVE')` is not present in the provided label.

- Exclusivity of XML tags highlighting an aspect term assigned to a combination of aspect category and sentiment polarity that is specified in the label. This includes ensuring that marked aspect terms with a particular combination do not occur more frequently in the generated text than specified in the label. Examples for valid and invalid generations are given in G.12.
- The generated example consisted of a single sentence, verified using the NLTK Tokenizer package.
- An XML tag must enclose text and must therefore not be empty.
- An aspect term, unlike an opinion term, must not exclusively consist of a single word from the following POS (Part-Of-Speech) classes: adjectives, verbs, conjunctions, determiners, interjections, and pronouns. Notably, it has also been considered that aspect terms comprising multiple words, including at least one of these POS classes, may be excluded. However, words from the mentioned POS classes can be part of an aspect term, for example, in the case of proper nouns (e.g., a restaurant named *"Little Goose"* includes the adjective *"Little"*). POS identification was employed using *spaCy* (*de_core_news_lg* model).

For a generated example, the presence of the other characteristics was examined only in case the first one was satisfied. In the case of an invalid XML scheme, it was unclear which part of the generation belonged to the tags or was part of the sentence and an aspect term. Consequently, the characteristics of these elements could not be evaluated.

If any of these characteristics were not met, the generation process was repeated with the same prompt. In the case that after 25 repeated generations no example was generated that met the aforementioned characteristics, new few-shot examples were selected in order to prevent an infinite loop. Particularly regarding GPT-3.5-turbo, where requests incurred costs, this could have become expensive over time.

In the context of LRS_{25}, the absence of further examples that could be used as few-shot examples led to a random rearrangement of the existing 25 examples within the prompt. In contrast, for LRS_{500}, 25 few-shot examples were again randomly selected from the 500 real examples.

Appendix H. Characteristics of generated examples leading to regeneration

See Table H.13.

Appendix I. Hyperparameter optimization for ACD and ACSA

Since there were only a limited number of examples available in both low-resource scenarios that could be used for hyperparameter optimization, the initially envisioned approach was to employ the hyperparameters used in related works for both tasks. For instance, in the study by Fehle et al. (2023), the best performing model on the ACD task was trained for 3 epochs, and on the ACSA task for 4 epochs. Similarly, in the work of Sun, Huang, and Qiu (2019), the ACD task was trained for 4 epochs.

However, during pre-experiments while implementing the model architecture using the dataset from Fehle et al. (2023), it was observed that for both the ACD and ACSA task, when using only 500 or 1,000 examples for training, the model was only predicting zero values during the first three epochs, requiring additional epochs for training.

In order to determine hyperparameters such as the number of epochs for the ACD and ACSA tasks for the sample sizes considered in conditions C1-C9 (500, 1,000, 1,500, and 2,000), the dataset by Fehle et al. (2023) was utilized. The characteristics of this dataset's examples closely align with those of the examples employed in the present study, since both comprise German-language examples, and they share an equivalent number of considered aspect categories (5) and sentiment polarities (3).

Similar to Fehle et al. (2023), *Optuna* (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) was employed for systematic hyperparameter optimization. 20 trials were conducted for each of the eight evaluation runs (since four sample sizes were considered for both the ACD and ACSA task), aiming to maximize the micro-averaged F1 score using a *Tree-structured Parzen Estimator* (TPE). A random subset of n training examples was selected from the 4,254 sentences of the dataset introduced by Fehle et al. (2023), alongside a random selection of 500 test examples. This process was repeated five times, which allowed for the evaluation with five different sets of training and test data in each trial for increased reliability.

- Learning rate $\in \{2e-5, 3e-5, 4e-5, 5e-5\}$
- Batch size $\in \{8, 16, 32\}$
- Number of epochs $\in [2, 20]$

The pre-selection of hyperparameters was based on the approach by Devlin (2018), with the search space of the number of epochs limited to a maximum of 20 epochs.

Table I.14 presents the hyperparameters of the best performing trial regarding the F1 score for both the ACD and ACSA tasks and the considered sample counts. As proceeded by Fehle et al. (2023) and Hellwig et al. (2024), a prediction was considered a true positive, if the predicted aspect(s) of a sentence (including the sentiment polarity for ACSA) occurred in the ground truth labels.

Table H.13
Characteristics of generated examples leading to regeneration.

	Llama-3-70B		GPT-3.5-turbo	
	LRS_{25}	LRS_{500}	LRS_{25}	LRS_{500}
# Examples	11,850	9,000	11,850	9,000
# Regenerations	12,656	4,411	901	258
# Examples with more than one Regeneration	2,308	1,664	260	108
# Examples with more than 25 Regenerations	92	19	9	0

(a) Regenerations

Characteristic Leading to a Regeneration	# Regenerations			
	Llama-3-70B		GPT-3.5-turbo	
	LRS_{25}	LRS_{500}	LRS_{25}	LRS_{500}
Invalid XML Schema	25	27	1	4
Invalid XML Tags (Invalid Attribute or Tag Name)	25	14	0	0
Aspect Category and Sentiment Polarity occur in text but not in label	1,147	887	93	31
Generated text is empty	0	0	0	0
More than one Sentence in Generated text	12,059	3,706	674	143
Empty Aspect Term in Generated text	1	0	0	0
Single Word Aspect Term with Specific Word Class[a]	997	359	179	86

(b) Characteristics leading to a regeneration

[a] Aspect terms consisting of one word are considered invalid, in case the word belongs to the following POS-classes: adjectives, verbs, conjunctions, determiners, interjections, and pronouns.

Table I.14
Best runs of hyperparameter optimization for ACD and ACSA.

Task	# Examples	# Epochs	Learning Rate	Batch Size	F1 Micro
ACD	500	13	3e−05	8	88.89
	1,000	14	2e−05	8	89.81
	1,500	10	3e−05	16	90.07
	2,000	7	3e−05	32	**90.11**
ACSA	500	15	2e−05	16	75.48
	1,000	15	2e−05	8	79.76
	1,500	20	2e−05	16	**81.06**
	2,000	18	3e−05	32	80.82

Appendix J. Preparation of data for the linguistic analysis

In order to calculate the metrics for real examples across the three sample sizes, we considered 1, 2 or 3 folds of 500 examples-subsets (see Fig. A.1(b) in the Appendix). The synthetic examples under consideration for this analysis of LRS_{500} correspond to the synthetic training examples considered in conditions C4, C5, and C6, comprising 500, 1,000, and 1,500 synthetic examples, respectively.

Since LRS_{25} did not require three sets of 500 synthetic examples each, but rather subsets of 475, 500, and 1,000 examples (as shown in Fig. A.1(c)), a different approach was adopted here. Firstly, the availability of three sets, each comprising 500 synthetic examples for analysis, had to be ensured. Secondly, within each set, every combination of aspect category and sentiment polarity specified in the labels had to occur with the same frequency as it is the case for the three sets of LRS_{500}.

To meet these requirements, the two sets comprising 500 and 1,000 synthetic examples were initially combined. As required, every combination of aspect category and sentiment polarity specified in the labels occurred with the same frequency in these two sets. Subsequently, the 1,500 examples were divided into three sets of 500 examples each using the stratification method for multi-label data proposed by Sechidis, Tsoumakas, and Vlahavas (2011). This ensured an even distribution of each combination of aspect category and sentiment polarity, as in a given set of 500 examples generated for LRS_{500}.

Data availability

Data will be made available on request.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631).

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, *55*(7), 1–39.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6788–6796).

Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2022). Survey of aspect-based sentiment analysis datasets. arXiv preprint arXiv:2204.05232.

Davis, E. (2024). Mathematics, word problems, common sense, and artificial intelligence. *American Mathematical Society. Bulletin*.

Dettmers, T., & Zettlemoyer, L. (2023). The case for 4-bit precision: k-bit inference scaling laws. In *International conference on machine learning* (pp. 7750–7774). PMLR.

Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023). Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews. In *Proceedings of the 19th conference on natural language processing (KONVENS 2023)* (pp. 202–218).

Fehle, J., Schmidt, T., & Wolff, C. (2021). Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques. In *Proceedings of the 17th conference on natural language processing (KONVENS 2021)* (pp. 86–103).

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications* (pp. 231–243). Springer.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Hellwig, N. C., Bink, M., Schmidt, T., Fehle, J., & Wolff, C. (2023). Transformer-based analysis of sentiment towards german political parties on Twitter during the 2021 election year. In *Proceedings of the 6th international conference on natural language and speech processing (ICNLSP 2023)* (pp. 84–98).

Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024). GERestaurant: A German dataset of annotated restaurant reviews for aspect-based sentiment analysis. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*.

Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019)* (pp. 34–41).

Li, K., Chen, C., Quan, X., Ling, Q., & Song, Y. (2020). Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7056–7066).

Li, G., Wang, H., Ding, Y., Zhou, K., & Yan, X. (2023). Data augmentation for aspect-based sentiment analysis. *International Journal of Machine Learning and Cybernetics*, *14*(1), 125–133.

Liesting, T., Frasincar, F., & Trușcă, M. M. (2021). Data augmentation in a hybrid approach for aspect-based sentiment analysis. In *Proceedings of the 36th annual ACM symposium on applied computing* (pp. 828–835).

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of natural language processing: Vol. 2*, (2010), (pp. 627–666). Oxfordshire.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on world wide web* (pp. 342–351).

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv preprint arXiv:2309.13638.

Meyer, S., Elsweiler, D., Ludwig, B., Fernandez-Pichel, M., & Losada, D. E. (2022). Do we still need human assessors? prompt-based GPT-3 user simulation in conversational ai. In *Proceedings of the 4th conference on conversational user interfaces* (pp. 1–6).

Møller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. arXiv preprint arXiv:2304.13861.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., et al. (2016). SemEval-2016 Task 5: Aspect based sentiment analysis. In *ProWorkshop on semantic evaluation (semEval-2016)* (pp. 19–30). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (semEval 2015)* (pp. 486–495).

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 27–35). Dublin, Ireland: Association for Computational Linguistics.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 1339–1384). Association for Computational Linguistics.

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (semEval-2017)* (pp. 502–518).

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, proceedings, part III 22* (pp. 145–158). Springer.

Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. arXiv:2310.18025.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2019). MASS: Masked sequence to sequence pre-training for language generation. In *International conference on machine learning* (pp. 5926–5936). PMLR.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT* (pp. 380–385).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Tripathy, A., Anand, A., & Rath, S. K. (2017). Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems*, *53*, 805–831.

Van Nooten, J., & Daelemans, W. (2023). Improving dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th workshop on computational approaches to subjectivity, sentiment, & social media analysis* (pp. 251–270).

Wang, A., Jiang, J., Ma, Y., Liu, A., & Okazaki, N. (2023). Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the the 12th joint conference on lexical and computational semantics (* SEM 2023)* (pp. 128–140).

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4195–4205).

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *Vol. 1*, In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*.

Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9209–9219).

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005.

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.