When Diverse Paraphrases Do Not Help: LLM Data Augmentation for Portuguese Aspect-Based Sentiment Analysis

Andrew B. Campos[1], Gabriel A. Gomes[1], Ulisses B. Corrêa[1], Larissa A. de Freitas[1]
[1]Universidade Federal de Pelotas (UFPel), Pelotas, Brazil
{abdcampos, gagomes, ulisses, larissa}@inf.ufpel.edu.br

Abstract

Data augmentation via LLM paraphrases is an appealing way to expand scarce training data for aspect-based tasks in low-resource languages. We test whether it helps on the Portuguese ABSAPT 2024 benchmark, using PTT5 as the target model. A simple PTT5 baseline already surpasses the best shared-task systems on both sub-tasks. We then augment training with paraphrases from two LLM families (Qwen2.5-7B and Gemma-3-Gaia-4B), generated at a high temperature with an anchoring prompt and selected by an explicit *diversity* filter based on low similarity to the source while preserving every aspect term verbatim. Although the retained paraphrases are measurably diverse and aspect-faithful by construction, augmentation yields no gain: aspect extraction consistently degrades across five seeds and both LLMs, while sentiment classification changes remain within seed variability. We characterise the generated data along semantic and surface-form axes and show that the injected diversity is predominantly lexical: the paraphrases move the training distribution away from the test set without providing evidence of additional semantic coverage beyond that of the human data.

1 Introduction

Data Augmentation (DA) is a technique that aims to increase training data while preserving the gold label. In this scenario, DA can be used for Natural Language Processing (NLP) tasks such as Aspect-Based Sentiment Analysis (ABSA) to improve model performance or achieve broader domain comprehension. ABSA is an NLP research area that aims to extract both the aspect term and its corresponding sentiment polarity in an opinion. These two sub-tasks are defined as Aspect Term Extraction (ATE) and Aspect Sentiment Classification (ASC). Figure 1 shows an example of ABSA.

The paraphrasing strategy has been widely employed for ABSA (Liesting et al., 2021; Wang et al., 2024). However, those approaches often result in complex frameworks. In this paper, we explore the direct use of Large Language Models (LLMs) for DA by generating paraphrases on the ABSAPT 2024 (Bender et al., 2024) dataset.

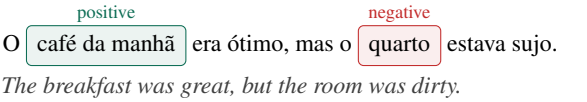


Figure 1: An ABSA-annotated hotel review with two aspect terms and their polarities.

The goal of our work is to elucidate the following question: does data augmentation through in-context learning with LLMs yield effective data to improve a model's performance? To explore this hypothesis, we train PTT5 (Carmo et al., 2020) on data augmented by two LLM families and evaluate the results over five seeds. Our contributions are threefold: (i) a simple PTT5 baseline that surpasses the best ABSAPT 2024 systems on both sub-tasks; (ii) a controlled negative result showing that diverse, aspect-anchored paraphrase augmentation does not improve over this baseline; and (iii) an analysis of the generated text indicating that diversity measured against the *source* is a poor proxy for usefulness, which tracks proximity to the *evaluation* distribution instead.

2 Related Work

Sobrevilla Cabezudo et al. (2024) explored the generation of abstract meaning representations through paraphrases. Their findings state that this approach can be a powerful DA strategy based on an analysis of the generated synthetic data. Their evaluation metrics encompass BLEU, METEOR, and Cosine Similarity to detect paraphrases with poor semantic relations.

Liesting et al. (2021) adapted EDA, back-translation, and word mixup specifically for the

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 119–123
October 17-21, 2026 ©2026 Association for Computational Linguistics

ABSA task, ensuring that target terms remained unchanged to preserve the sentence's original polarity. Their proposed modifications yielded positive performance gains on the SemEval 2015 and 2016 sentiment classification sub-tasks.

Ouyang et al. (2023) proposed the ABSA-ESA framework, which transforms implicit sentiment instances into explicit ones through text augmentation. Their approach leverages a T5 model for both data augmentation and fine-tuning. This method significantly improves implicit sentiment classification by enabling the model to accurately distinguish between explicit and implicit expressions.

The works cited previously yielded positive results when using paraphrasing for DA. However, this approach is often embedded within a framework that mixes encoders, decoders, and traditional DA methods like EDA. In this paper, we investigate whether a simple in-context learning approach can yield similar results in ABSA. Closest to our setting, the best ABSAPT 2024 systems were themselves trained on GPT-augmented data (Bender et al., 2024); we differ in starting from a baseline that surpasses them with human data alone, which isolates the contribution of the synthetic data.

3 Methodology

In this section, we discuss the empirical setup applied in this paper, starting with the dataset profile.

3.1 Dataset

We employed the official ABSAPT 2024 dataset, which contains hotel reviews extracted from the TripAdvisor website. The training set contains 1320 reviews with 4828 aspect annotations. The test set contains 283 reviews, providing 1321 gold aspect terms for the ATE sub-task and 1176 labelled aspect–polarity pairs for the ASC sub-task.

3.2 Paraphrase Generation

We generate paraphrases with two instruction-tuned LLM families to avoid model-specific conclusions: Qwen2.5-7B (Team, 2024) and Gemma-3-Gaia-4B (CAMILO-JUNIOR, 2025), a decoder model optimized for Brazilian Portuguese. Both were served using vLLM (Kwon et al., 2023) on a single RTX 5090 (32 GB).

For each training review, we generate four paraphrases at a temperature of 1.5 (remaining decoding parameters at their defaults) to maximize lexical divergence, using a chat prompt (Appendix A) with three fixed, manually written demonstrations that *anchor* the aspect terms: the model is instructed to copy every aspect term verbatim while rewriting the remaining context. Each paraphrase inherits the source annotations; paraphrases that do not reproduce every aspect term verbatim are discarded, ensuring no label is left unanchored. Even at temperature 1.5, anchoring succeeds for 78–80% (Qwen2.5-7B) and 73–76% (Gaia-4B) of the raw generations, leaving at least one fully anchored paraphrase for over 93% of the training reviews.

Generation is repeated under five seeds (42, 117, 584, 2026, and 9999), which control the sampling of the LLM; the same seeds later control training stochasticity, so each seed indexes one end-to-end run from generation to evaluation.

3.3 Data Selection

We retain only diverse paraphrases, defined as those exhibiting low surface similarity and strong lexical rewriting relative to the source text. Degenerate outputs (empty, identical to the source, or identical to a test review) are removed first. For each remaining paraphrase p of a source review t, we compute the cosine similarity between their sentence embeddings:

$$\cos(t, p) = \frac{\mathbf{e}_t \cdot \mathbf{e}_p}{\|\mathbf{e}_t\| \, \|\mathbf{e}_p\|}, \tag{1}$$

where $\mathbf{e}$ are produced by a Portuguese sentence encoder,[1], a BERTimbau-based (Souza et al., 2020) Sentence-Transformers model (Reimers and Gurevych, 2019) fine-tuned for semantic textual similarity and the sentence-level sacreBLEU score (Post, 2018) of p against t. A paraphrase is kept only when:

$$\cos(t, p) < 0.85 \quad \text{and} \quad \mathrm{BLEU}(p, t) < 30, \tag{2}$$

and provided all its original aspect terms are present. The BLEU and cosine thresholds enforce substantial lexical and representational divergence from the source, while the anchoring constraint maintains aspect-label validity. Out of the 5280 candidates generated per seed (4×1320), the full pipeline retains 1193 (Qwen2.5-7B, 23%) and 1977 (Gaia-4B, 37%) paraphrases on average. Table 1 characterizes the retained data: the selected paraphrases are heavily rewritten (BLEU ≈ 10) yet remain semantically related to the source (cosine ≈ 0.79).

[1] rufimelo/bert-large-portuguese-cased-sts

Paraphraser	Kept/seed	Cosine	BLEU
Qwen2.5-7B	1193 ± 8	0.79 ± 0.07	11.2 ± 6.0
Gaia-4B	1977 ± 30	0.79 ± 0.07	9.3 ± 5.4

Table 1: Retained diverse paraphrases over five seeds: number kept per seed, and their cosine similarity and BLEU to the source review.

Training data	ATE (F1)	ASC (BAcc)
Baseline	0.832 ± 0.003	0.775 ± 0.008
+ Qwen2.5-7B	0.816 ± 0.009	0.767 ± 0.008
+ Gaia-4B	0.821 ± 0.009	0.766 ± 0.010

Table 2: Diverse paraphrase augmentation on both tasks, mean $\pm$ standard deviation over five seeds.

Corpus	N	Cosine	BLEU
Human (original)	1317	0.752	5.81
Qwen2.5-7B	1193	0.725 ± 0.001	4.43 ± 0.08
Gaia-4B	1977	0.730 ± 0.001	4.30 ± 0.03

Table 3: Proximity of each training corpus to the human test set. N counts unique texts after normalisation (three near-duplicate human reviews merge); the human corpus is fixed, while the synthetic corpora vary by seed (mean $\pm$ standard deviation).

3.4 Model

We train a Portuguese T5 (Carmo et al., 2020) model (PTT5 Base, 220M parameters) in a text-to-text setup. For ATE, the input is the raw review and the target is the list of its gold aspect terms, sorted and joined by a separator; predictions are parsed back into a set and scored with micro F1 under exact string match after case and whitespace normalization.

For ASC, the input concatenates the aspect term and the review and the target is the polarity label, scored with balanced accuracy (macro-averaged recall). In augmented configurations, the training set is the union of the original data and the retained paraphrases. The model was trained for 10 epochs with the AdamW optimizer, a learning rate of 3×10^{-4}, and a batch size of 16; hyperparameters were fixed a priori, without a development set, and we evaluate the final checkpoint. Each configuration was run with the same five seeds: the seed controls training stochasticity and, in augmented configurations, also selects the corresponding paraphrase set. Results are reported as mean $\pm$ standard deviation.

4 Results and Discussion

Before turning to augmentation, we establish the capacity of the target model on the original data. Averaged over five seeds, PTT5 Base reaches an F1 of 0.832 ± 0.003 on aspect term extraction and a balanced accuracy of 0.775 ± 0.008 on aspect sentiment classification. Both figures surpass the best systems reported by Bender et al. (2024), which reached 0.637 F1 and 0.652 BAcc using a BERTimbau (Souza et al., 2020) Large model trained on GPT-augmented data; the comparison uses the official test set and metrics of the shared task. Since PTT5 Base is considerably smaller than BERTimbau Large, we attribute the gap primarily to the text-to-text formulation rather than to model capacity. The per-seed spread is narrow—the extraction scores fall between 0.827 and 0.836, and the classification scores between 0.761 and 0.781—so the seed-level variation is an order of magnitude smaller than the gap to the shared-task systems. The baseline is therefore strong and stable, which is the setting in which additional synthetic data has the least room to help.

The experimental outcomes for all seeds on the augmented training set are detailed in Table 2. Diverse, aspect-preserving augmentation does not yield improvements for either sub-task: both LLM setups land below the baseline on aspect extraction and aspect sentiment classification. A paired t-test over seeds finds the extraction drop significant for Qwen2.5-7B ($p = 0.045$) and consistent in sign but not significant for Gaia-4B ($p = 0.11$); the classification differences lie within seed variability ($p \geq 0.29$). The paired delta is negative for four of the five seeds in every configuration, and no configuration improves over the baseline on either sub-task.

4.1 How close is the data to the test set?

To understand why diverse paraphrases failed to help, we measure how close each training corpus sits to the human *test* set, separating the semantic axis (embedding cosine) from the surface, syntactic axis (BLEU). We reuse the encoder and BLEU configuration of the diversity filter: for every training text, we report its maximum cosine similarity over the test reviews and its maximum sentence-level BLEU among its ten nearest test neighbours by cosine, averaged over the corpus (Table 3). Since the train and test sets contain disjoint reviews, the absolute BLEU is low by construction; the comparison is meaningful primarily *across* the three corpora.

On both evaluation axes, the original human data remains the closest to the test set (cosine 0.752, BLEU 5.81). Paraphrasing systematically moves the training distribution *away* from the test targets, doing so asymmetrically: the semantic drop is small (-0.02 to -0.03 cosine, $\approx$ 3–4%) while the surface syntactic drop is proportionally much larger (BLEU decreases by -24 to -26%). The injected diversity is thus mostly lexical, and it introduces variations that are *not* aligned with the test set's surface forms. Crucially, the paraphrases provide no evidence of additional semantic coverage aligned with the test distribution the human data lacked—the originals are semantically *closer* to the test set than either automated paraphraser—while distorting the surface statistics of the training set. This is consistent with the minor performance drop below the baseline shown in Table 2: the target model is not starved of the semantics it is evaluated on, so the extra syntactic variety would dilute rather than enrich the core training signal.

5 Conclusions

In this paper, we showed that augmenting training data with diverse, LLM-generated paraphrases fails to improve performance on a high-baseline Portuguese ABSA task. Our findings highlight two critical insights for synthetic data generation in aspect-based tasks.

First, when the original training set already offers comprehensive semantic coverage aligned with the test distribution, introducing heavily rewritten paraphrases can shift the surface statistics and inject lexical noise rather than useful signal. This suggests that lexical diversity alone, without semantic expansion, dilutes the training efficiency of the target model. It also indicates that diversity metrics computed against the source text are a poor proxy for the usefulness of generated data: what matters is the distance to the evaluation distribution, not the distance to the seed.

Second, while engineering heuristics—such as prompt anchoring, temperature tuning, and strict filtering—can preserve labels and ensure semantic proximity, this process introduces a trade-off: the calibration effort these heuristics require works against the primary appeal of LLMs for automated, low-resource data scaling.

Ultimately, our results show that PTT5 robustly handles the presence of synthetic data, experiencing only a minor performance degradation. This indicates that while the model is resilient to distribution shifts, standard in-context paraphrasing is insufficient to push performance past an already strong human-data baseline.

Limitations

Our study is limited to one Portuguese hotel-review corpus and to PTT5 Base as the target model; the conclusions may not transfer to other domains or architectures. Hyperparameters were fixed a priori, without a development set. We test a single augmentation policy—appending all retained paraphrases to the original data—so other augmentation ratios or mixing strategies might behave differently. With five seeds and a fixed split, only the Qwen2.5-7B extraction drop is individually significant; our claim is therefore the *absence of gains* above roughly one F1 point, not a uniformly significant degradation. We study two open-weight LLM paraphrasers; results may differ for larger or closed API models. Finally, our proximity analysis relies on a single sentence encoder and on sentence-level BLEU, both of which are imperfect proxies for semantic and surface similarity.

Supplementary Material Statement

The full paraphrasing prompt, including its three fixed demonstrations, is given in Appendix A. The code for paraphrase generation, filtering, and training, together with the generated paraphrases for all five seeds can be found here https://github.com/AndrewwBC/INLG-2026. All models used (PTT5 Base, Qwen2.5-7B, Gemma-3-Gaia-4B) are publicly available and were used in accordance with their licences.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors thank the reviewers for their valuable contributions.

References

Alexandre Thurow Bender, Gabriel A Gomes, Emerson P Lopes, Ricardo M Araujo, Larissa A de Freitas, and Ulisses B Corrêa. 2024. Overview of absapt at iberlef 2024: Overview of the task on aspect-based sentiment analysis in portuguese. *Procesamiento del Lenguaje Natural*, 73:315–322.

S. S. T.; PEREIRA L. A.; AMADEUS M.; SCOTTI R.; FAZZIONI D.; NOVAIS A. M. A.; JORDÃO

S. A. A. CAMILO-JUNIOR, C. G.; OLIVEIRA. 2025. Gaia: An open language model for brazilian portuguese. https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Tomas Liesting, Flavius Frasincar, and Maria Mihaela Truşcă. 2021. Data augmentation in a hybrid approach for aspect-based sentiment analysis. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 828–835, New York, NY, USA. Association for Computing Machinery.

Jihong Ouyang, Zhiyao Yang, Silong Liang, Bing Wang, Yimeng Wang, and Ximing Li. 2023. Aspect-based sentiment analysis with explicit sentiment augmentations. *Preprint*, arXiv:2312.10961.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo, Marcio Lima Inacio, and Thiago Alexandre Salgueiro Pardo. 2024. Investigating paraphrase generation as a data augmentation strategy for low-resource AMR-to-text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 663–675, Tokyo, Japan. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Haining Wang, Kang He, Bobo Li, Lei Chen, Fei Li, Xu Han, Chong Teng, and Donghong Ji. 2024. Refining and synthesis: A simple yet effective data augmentation framework for cross-domain aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10318–10329, Bangkok, Thailand. Association for Computational Linguistics.

A Paraphrasing Prompt

To ensure complete reproducibility of the data generation process, we provide the exact chat prompt used to instruct both Qwen2.5-7B and Gemma-3-Gaia-4B. The prompt was executed in Portuguese to prevent cross-lingual interference. The system message is:

```
Você reescreve avaliações de hotel em
↪ português com OUTRAS palavras e
↪ OUTRA construção, preservando o
↪ sentido.

Você recebe o texto e uma lista de
↪ TERMOS DE ASPECTO. Regras
↪ obrigatórias:
- Cada termo de aspecto aparece
↪ literalmente no texto e DEVE ser
↪ copiado na paráfrase com a MESMA
↪ grafia, sem tradução, flexão ou
↪ sinônimo.
- Reescreva livremente todo o resto
↪ (opiniões, conectivos, estrutura).
- Não invente aspectos novos nem
↪ remova os existentes.
- Devolva APENAS o texto reescrito,
↪ sem comentários.

- $List of few-shots$.
```