Automated Topic Annotation in Brazilian Product Reviews: A Case Study of Adversarial Examples with Sabia-3

Lucas Nildaimon dos Santos Silva[1], **Livy Real**[2]

[1] Department of Computing, Federal University of São Carlos, Brazil

[2]Federal University of Paraná, Brazil

lucas.silva@estudante.ufscar.br, livy.real@gmail.com

Abstract. *High-quality annotated data is essential for many Natural Language Processing tasks, but traditional human annotation methods are often resource-intensive. Large Language Models (LLMs) offer potential solutions by generating labels for training datasets. This paper explores the effectiveness of using the Sabiá-3 LLM for automatically labeling data for a multi-label topic classification task in Brazilian Portuguese product reviews. We compare the performance of Sabia-3-generated labels against human annotations using the RePro dataset. The study evaluates Sabiá-3 on both random and adversarial datasets, highlighting its strengths in frequent topics, while identifying limitations in more nuanced categories. Models trained on Sabiá-3 annotations showed promising results in common categories but faced challenges with ambiguous cases. Our findings suggest that while LLMs can streamline parts of the annotation process, human oversight remains essential, particularly in complex or less frequent cases. This research contributes new insights into the use of LLMs for automated data annotation in Brazilian Portuguese.*

1. Introduction

High-quality annotated data has long been a critical challenge in Natural Language Processing (NLP). Traditional human annotation is both time-consuming and costly, often requiring specialized knowledge that can be difficult to obtain, particularly in niche fields such as legal, medical, or industrial sectors like oil and gas. These constraints make acquiring high-quality data challenging in both academic research and resource-constrained industries.

With the advent of Large Language Models (LLMs), many traditional NLP tasks are being reevaluated. LLMs are increasingly applied to end-to-end tasks, particularly those involving creativity or text generation, such as conversational agents. However, for many downstream tasks—like spam detection, document classification, and sentiment analysis—that already benefit from classical algorithms when trained on appropriate data, substituting LLMs may not be necessary or practical.

Instead, for many downstream tasks not related to generative applications, it is often more advantageous to use LLMs to generate data for training classical models rather than replacing established NLP pipelines with LLMs. Nevertheless, relying on LLM-generated labels poses risks, especially in the absence of a human-labeled test set for validation. The state-of-the-art performance of various models across different languages

and domains is not fully understood, and LLM capabilities are rapidly evolving, making current benchmarks potentially obsolete in the near future.

This paper contributes to this ongoing discussion by focusing on a specific application: multi-label topic classification in Product Reviews in Brazilian Portuguese. We compare human-annotated data with LLM-annotated data and assess the performance of models trained on these datasets. Our approach includes a method for balancing random samples with an adversarial dataset.

Product reviews, crucial in the e-commerce and marketplace sectors, significantly influence consumer purchasing decisions. Given the availability of review datasets in Brazilian Portuguese, our findings are not only relevant to this industry but also applicable to other contexts.

Our study focuses on evaluating the Sabiá-3 LLM [Almeida et al. 2024] for the annotation of a multi-label topic classification task in Brazilian Portuguese product reviews. We use the RePro dataset [dos Santos Silva et al. 2024] and compare the performance of Sabiá-3 with human annotators on both general and complex cases.

The key question driving this study is whether Sabiá-3, given the same guidelines as human annotators, can perform topic labeling with comparable accuracy and consistency. By comparing its performance against human annotations on both general and challenging adversarial datasets, we aim to provide insights into its viability as a replacement for human annotators in this specific task.

Our objectives were threefold: to compare human and LLM-generated annotations, to evaluate Sabiá-3's performance on both random and adversarial datasets, and to assess the quality of models trained on human-labeled and LLM-labeled data.

2. Related Works

We investigated two primary areas: prominent datasets of Product Reviews in Portuguese and recent advancements in using Large Language Models (LLMs) for data annotation.

Regarding Product Reviews, a well-established textual genre on the web [Pollach 2006], multiple datasets are available in Brazilian Portuguese. The earliest is the 'Brazilian E-Commerce Public Dataset by Olist'[1], released in 2018. This dataset encompasses approximately 100,000 orders from 2016 to 2018, including details on order status, pricing, product attributes, and customer reviews. In 2019, [Real et al. 2019] introduced the B2W-Reviews-01 dataset, which contains over 130,000 product reviews and includes additional information such as reviewers' gender, age, and location, along with product evaluations like a 5-star rating and a "recommend-to-a-friend" question answered by all reviewers.

Several studies have built upon B2W-Reviews-01. [Real et al. 2020] conducted the first analysis of topics within product reviews, while Brands.Br[2] [Fonseca et al. 2020] incorporated brand information to fill gaps in B2W-Reviews-01. [Zagatti et al. 2021] focused on anonymizing the B2W-Reviews-01 corpus to ensure compliance with the General Data Protection Law.

[1]https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
[2]https://github.com/metalmorphy/Brands.Br

[dos Santos Silva et al. 2024] offers a comprehensive examination of topics in product reviews, extending the work of [Real et al. 2020]. It introduces the RePro corpus, a 10,000-sample subset of B2W-Reviews-01, annotated with topics by human experts. This corpus is available for non-commercial use on GitHub[3] and HuggingFace[4] under the CC BY-NC-SA 4.0 license.

Regarding LLMs for data annotation, [Ye et al. 2022] introduced ZEROGEN, enhancing zero-shot learning by generating task-specific datasets with LLMs for efficient inference. [Ding et al. 2023] evaluated GPT-3's annotation performance, comparing it with traditional methods across various NLP tasks.

In contrast to previous works, our study focuses on evaluating the Sabiá-3 model for data annotation in Brazilian Portuguese, using a real-world dataset (RePro). While prior research explores LLMs like GPT-3 and GPT-4, often in English and for zero-shot or synthetic data generation, we provide a direct comparison between human and LLM annotations in a less-studied language. This contributes new insights into the effectiveness of LLMs for data annotation in non-English contexts.

3. Methodology

We selected Sabiá-3 for this study as it is the latest Brazilian LLM specifically trained on Brazilian Portuguese data. Its predecessor, Sabiá-2, was evaluated across 64 diverse exams, ranging from university entrance exams to professional certification and graduate-level tests, where it outperformed GPT-4 in 23 of the 64 assessments [Almeida et al. 2024]. Additionally, [Inacio and Oliveira 2024] demonstrated Sabiá-2's effectiveness in humor generation, showing it to be on par with rule-based approaches for this task. By evaluating Sabiá-3, we aim to contribute to the growing body of research on Brazilian LLMs and further explore their capabilities.

To conduct our experiments, we selected two subsamples from RePro: a random sample representing general cases, and an adversarial dataset [5] where human annotators disagreed on the assigned topics. For both samples, we re-annotated the reviews with Sabiá-3, following the original topic annotation guidelines. We then trained two models: one using human-labeled data and another using LLM-generated labels. Both models were evaluated against human-labeled test data to assess the quality of the annotations and the effectiveness of the trained models.

3.1. Dataset

We used the annotated samples from the RePro dataset, which consists of product reviews in Brazilian Portuguese. Each review in RePro is annotated with one or more of the following six topics:

- **ANÚNCIO (Advertisement):** Contexts where the delivered product aligns or misaligns with the information presented on the product's webpage, such as description, images, technical specifications, and overall advertising.
- **PRODUTO (Product):** Comments on product quality, originality, value for money, attributes, user experience, and general compliments.

[3]https://github.com/lucasnil/repro

[4]https://huggingface.co/datasets/lucasnil/repro

[5]Available at: https://github.com/lucasnil/repro

- **ENTREGA (Delivery):** Related to the speed of delivery, time, non-delivery, in-store pick-up, virtual delivery (e.g., gift cards, codes), and freight comments.
- **CONDIÇÕES DE RECEBIMENTO (Receipt Conditions):** Comments about the state of the product upon receipt, such as damage, packaging quality, incomplete or incorrect orders, and whether the product met the customer's expectations.
- **OUTROS (Others):** Contexts involving seller inquiries, customer service, stock availability, shopping experience, payment methods, or nonsensical information that is not harmful to the company.
- **INADEQUADA (Inadequate):** Harmful information, such as profanity, competitor mentions, legal references, external links, or personal information.

Text	Topic
Adorei, A caixa veio bem amassada mas o produto estava em ordem.. já usei e gostei	CONDICOESDERECEBIMENTO, PRODUTO
I loved it, the box came quite dented, but the product was in good condition.. I've already used it and liked it	*RECEIVINGCONDITIONS, PRODUCT*
A cor desse celular não é dourado igual da imagem da foto, ele é beeeem mais claro!!	ANUNCIO
The color of this phone is not gold like in the picture, it's muuuch lighter!!	*ADVERTISEMENT*

Table 1. Examples of RePro

Two distinct subsamples of the RePro dataset were used in the LLM annotation experiment:

- **Random Sample:** A random selection of 1400 reviews from the RePro dataset, providing a broad spectrum of product feedback.
- **Adversarial Sample:** A subset of 2454 reviews in which the original human annotators disagreed on the assigned topic labels, requiring a third annotator to resolve the conflicts. We hypothesize that this makes the dataset more challenging for automatic labeling.

3.2. LLM Annotation Promt

The automatic annotation process was conducted using the Sabiá-3 LLM, which was tasked with assigning one or more of the six predefined topics to each review. The prompt used for the labeling task (translated from Portuguese) was as follows:

> *"You are an automatic product review labeler for an e-commerce platform. You must read and label reviews with one or more of the following six topics: ANÚNCIO, PRODUTO, ENTREGA, CONDIÇÕES DE RECEBIMENTO, OUTROS, and INADEQUADA. To do this, you must strictly follow the annotation guidelines provided."*

The annotation guidelines, included with the prompt message, correspond to the topic descriptions outlined in Section 3.1. Each review was presented to the model via this prompt, and the Sabiá-3 LLM generated responses. The Sabiá-3 model was integrated through the Langchain framework, utilizing the Sábia-3 API to carry out the labeling task.

3.3. Human vs. LLM-Generated Labels

To assess the impact of label quality, we fine-tuned two BERT models for multilabel classification: one using human-annotated labels and another using labels generated by the Sabiá-3 LLM. Both models were trained on the Random Sample dataset and evaluated on a separate test set of 1000 human-labeled samples.

We employed BERTimbau, a pre-trained BERT model for Brazilian Portuguese. The models were fine-tuned using standard hyperparameters optimized for multilabel classification. Training was conducted for 3 epochs with a learning rate of 2e-5, a batch size of 16 (training) or 64 (evaluation), and a weight decay of 0.01. Model performance was monitored every 10 steps, with the best-performing model (based on F1 score) saved to ensure optimal balance between precision and recall across labels.

4. Results

In this section, we present the results achieved by Sabiá-3 on the task of automatically labeling product reviews with multilabel topics and the comparison of models trained on human and LLM-generated labels.

4.1. Automatic Labeling

The results for the Random Sample dataset and the Adversarial dataset are displayed in Table 2 and 3, respectively. We evaluate the LLM using accuracy, precision, and F1-score for each class.

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.62	0.83	0.71	121
ENTREGA	0.93	0.98	0.96	455
PRODUTO	0.94	0.96	0.95	1087
CONDICOES DE RECEBIMENTO	0.55	0.80	0.65	229
INADEQUADA	0.25	0.45	0.32	58
OUTROS	0.86	0.41	0.56	344
Macro average	0.69	0.74	0.69	2294

Table 2. Performance metrics for the Random Sample dataset

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.65	0.84	0.73	429
ENTREGA	0.87	0.96	0.91	947
PRODUTO	0.82	0.90	0.86	1566
CONDICOES DE RECEBIMENTO	0.69	0.74	0.71	822
INADEQUADA	0.34	0.44	0.38	231
OUTROS	0.87	0.36	0.51	1118
Macro avg	0.71	0.71	0.68	5113

Table 3. Performance metrics for the Adversarial dataset

Across both datasets, the Sabiá-3 model demonstrated strong performance on the ENTREGA and PRODUTO classes, which consistently achieved high precision and recall values. These results highlight the model's effectiveness in identifying frequent and well-defined topics. However, the model struggled significantly with the INADEQUADA and OUTROS classes, where both precision and recall were notably lower than for other classes. For INADEQUADA, the model exhibited very low precision, indicating a high rate of false positives, meaning that it frequently mislabeled reviews as inappropriate

when they were not. This could suggest that the model is overly sensitive to certain keywords or language patterns that it associates with harmful content, even in cases where human annotators would not. The recall for this class was also low, reflecting the model's difficulty in identifying a substantial portion of truly inappropriate reviews. Similarly, for OUTROS, while the precision was relatively high, the recall was much lower, which implies that the model was conservative in assigning this label. It only identified a subset of the true instances of this class but was generally accurate when it did. This suggests the model may be underrepresenting miscellaneous topics that don't fit cleanly into the other predefined categories.

4.2. Comparison of Models Trained on Human vs. LLM-Generated Labels

The results for both models, with human and LLM labels, are displayed in Tables 4 and 5, respectively.

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.97	0.51	0.67	72
ENTREGA	0.96	0.97	0.97	317
PRODUTO	0.96	0.94	0.95	774
CONDICOES DE RECEBIMENTO	0.91	0.69	0.79	180
INADEQUADA	0.00	0.00	0.00	34
OUTROS	0.85	0.64	0.73	235
Macro avg	0.78	0.63	0.68	1612

Table 4. Classification performance metrics for the model tuned with human labels

Class	Precision	Recall	F1-Score	Support
ANUNCIO	0.67	0.78	0.72	72
ENTREGA	0.95	0.97	0.96	317
PRODUTO	0.93	0.95	0.94	774
CONDICOES DE RECEBIMENTO	0.63	0.69	0.66	180
INADEQUADA	1.00	0.03	0.06	34
OUTROS	0.91	0.22	0.35	235
Macro avg	0.85	0.61	0.62	1612

Table 5. Classification performance metrics for the model trained with LLM labels

The model trained on human labels exhibited more balanced performance across most classes, particularly excelling in recall, indicating it identified more relevant instances. However, it struggled with the INADEQUADA class, which had the lowest metrics possible. In this case, this should be due to fewer training examples of this particular class. However, previous work, such as in [dos Santos Silva et al. 2024], showed that class INADEQUADA continued to perform poorly despite a larger number of samples, suggesting inherent challenges in labeling this class accurately, likely due to its ambiguity.

The model trained on LLM-generated labels demonstrated greater variability across classes. It maintained strong performance in common categories like PRODUTO and ENTREGA, but its precision for the OUTROS class was notably lower compared to

the human-labeled model, suggesting difficulty in distinguishing this category. In contrast, the LLM-generated labels yielded higher recall for the ANUNCIO class, indicating the model identified more instances but at the cost of precision, likely including more irrelevant cases. As with the human-labeled model, performance for the INADEQUADA class remained low.

5. Automatic Labeling Qualitative Analysis

In analyzing the quality of the Sabia-3 outputs, the most noticeable issue was the occurrence of hallucinations—69 cases out of 3,854 instances—and how these hallucinations manifested. We define hallucinations in two ways: when the model introduces a new topic not covered in the guidelines, and when it provides an explanation for a label.

Although we did not encounter any output that was entirely misaligned with the task, Sabia-3 often attempted to be more specific than necessary.

Consider the following example: *O produto é bom nos primeiros 6 meses, depois começa a dar problemas. Esse é o segundo que compro, pois o primeiro tive o dinheiro ressarcido, aí comprei este. Recomendo comprar sempre com a garantia extendida. Guardem a nota fiscal do produto e da garantia extendida. É funcional pois serve também para vigiar a casa e os pets.*[6].

A human labeled this as: *OUTROS, PRODUTO*, since the review discusses the product, and warranty was explicitly categorized under the *OUTROS* topic in the guidelines. Sabia-3, however, produced the following output: *PRODUTO, GARANTIA (dentro de OUTROS), CONDIÇÕES DE RECEBIMENTO (referente à garantia e ressarcimento)*. While the model correctly labeled *PRODUTO*, it tried to be more specific by introducing a new label, *GARANTIA* (warranty), though it accurately recognized this as part of the *OUTROS* topic. The confusion arose with *CONDIÇÕES DE RECEBIMENTO*, which is a label intended to describe the state of the product upon receipt, not conditions after the product has been received.

Sabia-3 also generated more specific, albeit incorrect, labels such as *CUSTO-BENEFÍCIO, ATENDIMENTO, ESTOQUE, EXPERIÊNCIA DE COMPRA, ATENDIMENTO AO CONSUMIDOR*.[7]. Interestingly, all of these topics are closely related to the review content and relevant to the domain, suggesting these are domain-related hallucinations rather than out-of-scope hallucinations. However, despite their relevance, these labels were incorrect for the task at hand and would require careful post-processing. It's worth noting that domain-related hallucinations are significantly harder to detect than those that are completely unrelated to the task.

Lastly, consider the review: *EXCELENTE COMPRA Comprei esse motor e adaptei na vassoura elétrica. A patroa adorou (e eu também). Ligou o bicho, saiu voando por aí e até hoje não voltou.*[8].

[6] *The product is good for the first 6 months, but then it starts having issues. This is the second one I've bought because I got a refund for the first one, so I purchased this one. I recommend always buying it with an extended warranty. Keep the receipt for both the product and the extended warranty. It's functional as it can also be used to monitor the house and pets.*

[7] *COST-BENEFIT, SUPPORT, STOCK, PURCHASE EXPERIENCE, CUSTOMER SERVICE*

[8] *EXCELLENT PURCHASE I bought this motor and adapted it to an electric broom. The wife loved it (and so did I). She turned it on, took off flying, and hasn't come back since.*

In this case, the model misclassified the topics but subtly acknowledged the humorous tone of the review, labeling it as *OUTROS (devido ao tom jocoso e à experiência do usuário)*[9], indicating an attempt to account for the review's humorous content.

6. Conclusions

In this study, we evaluated the robustness of Sabiá-3, a Brazilian Large Language Model (LLM), in the task of multi-label topic annotation for product reviews. By comparing its performance against human-labeled data, we investigated both general cases from a random sample and more challenging cases from an adversarial sample in the RePro dataset. Additionally, we trained models on both human and LLM-generated labels to assess the impact of label quality on downstream model performance.

Our findings indicate that Sabiá-3 performs well in identifying clear and frequent topics, such as ENTREGA and PRODUTO, but faces significant challenges in more ambiguous categories like INADEQUADA and OUTROS. These discrepancies underscore the limitations of LLMs when tasked with handling nuanced or infrequent cases, which require a more sophisticated understanding of context.

From the qualitative analysis, we observed that Sabia-3 occasionally produced hallucinations, introducing labels not covered in the original guidelines or over-specifying topics. While these hallucinations were often domain-relevant, they deviated from the task's specific requirements, indicating the model's tendency to infer context too aggressively. This over-specification led to issues in cases where the model introduced new labels like GARANTIA or misapplied existing ones, such as using CONDIÇÕES DE RECEBIMENTO inappropriately. These domain-related hallucinations are particularly concerning because they are harder to detect than out-of-scope errors, necessitating careful post-processing when relying on LLM outputs.

The comparison between models trained on human-labeled and LLM-labeled data further highlighted the challenges posed by ambiguous cases. While LLM-generated labels performed adequately in simpler categories, the human-labeled models provided a more balanced and accurate representation across all categories, particularly for complex or less frequent classes. This emphasizes the continued importance of human oversight in training data for high-stakes NLP tasks.

In conclusion, while LLMs like Sabiá-3 show promise in automating parts of the annotation process, particularly for well-defined and frequent topics, they struggle with edge cases and can introduce misleading labels. Our study contributes to the ongoing evaluation of LLMs, demonstrating the importance of including both general and adversarial datasets to test their limits. The hallucination issue also highlights the need for improved LLM interpretability and error correction mechanisms.

Future research should focus on developing advanced techniques for the automatic detection of adversarial examples within general datasets. Such methods could help streamline the annotation process by identifying challenging or ambiguous cases that typically require human intervention. This would not only reduce the reliance on human effort but also enhance the overall quality and reliability of LLM-generated annotations, enabling more efficient and accurate handling of complex tasks.

[9] *OTHERS (due to the playful tone and the user's experience).*

References

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models.

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., and Bing, L. (2023). Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

dos Santos Silva, L. N., Real, L., Zandavalle, A. C. B., Rodrigues, C. F. G., da Silva Gama, T., Souza, F. G., and Zaidan, P. D. S. (2024). RePro: a benchmark for opinion mining for Brazilian Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 432–440, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Fonseca, E., Oliveira, A., Gadelha, C., and Guandaline, V. (2020). Brands.br - a portuguese reviews corpus. In *OpenCor*.

Inacio, M. L. and Oliveira, H. G. (2024). Generation of punning riddles in portuguese with prompt chaining paper type: Late breaking results. *15th International Conference on Computational Creativity (ICCC'24)*.

Pollach, I. (2006). Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 51c–51c.

Real, L., Bento, A., Soares, K., Oshiro, M., and Mafra, A. (2020). B2w-reviews02, an annotated review sample. In *OpenCor*.

Real, L., Oshiro, M., and Mafra, A. (2019). B2w-reviews01-an open product reviews corpus. In *the Proceedings of the XII Symposium in Information and Human Language Technology*, pages 200–208.

Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., and Kong, L. (2022). Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669.

Zagatti, F., Silva, L., and Real, L. (2021). Anonymization of the b2w-reviews01 corpus. In *OpenCor*.