RePro: A Benchmark Dataset for Opinion Mining in Brazilian Portuguese

Lucas Nildaimon dos Santos Silva [1], **Ana Cláudia Zandavalle** [2]
Carolina Francisco Gadelha Rodrigues [3], **Tatiana da Silva Gama** [3]
Fernando Guedes Souza [4], **Phillipe Derwich Silva Zaidan** [4]
Alice Florencio Severino da Silva [5], **Karina Soares** [6], **Livy Real** [7]
[1] Department of Computing, Federal University of São Carlos, Brazil
[2] Federal University of Santa Catarina, Florianópolis, Brazil
[3] Americanas S.A., Rio de Janeiro, Brazil
[4] Federal University of Mina Gerais, Belo Horizonte, Brazil
[5] Getúlio Vargas Foundation, Rio de Janeiro, Brazil
[6] Mutant, São Paulo, Brazil
[7] Quinto Andar Inc, São Paulo, Brazil
lucas.silva@estudante.ufscar.br
tatiana.gama@americanas.io
karinasoares@tuta.io
{ana.zandavalle,carolfgr25,f.guedes93,alice.florencio,livyreal}@gmail.com

Abstract

We introduce RePro, a corpus of e-commerce product reviews in Brazilian Portuguese labeled with sentiment and topic information. We carried out a careful annotation process, whose aim is to introduce an easily available and open benchmark for opinion mining related tasks, namely sentiment analysis and topic modeling tasks. This work describes the corpus design and annotation process as well as the preliminary results of classification tasks. These preliminary results can be used as baselines for future work. RePro contains 10,000 humanly annotated reviews, based on data from the largest Brazilian e-commerce platform, which produced the B2W-Reviews01 dataset.

1 Introduction

The availability of open, plentiful and high-quality data is still one of the main bottlenecks of Natural Language Processing. When it comes to low-resource languages, such as Brazilian Portuguese, this challenge is even bigger. This work introduces the RePro (REview of PROducts) corpus, a humanly annotated sample of the large B2W-Reviews-01 corpus (Real et al., 2019) containing 10,000 samples annotated with sentiment and topic information. With RePro, we aim to offer to the NLP community, a benchmark for tasks related to opinion mining, namely sentiment analysis (SA) and topic modeling (TM). We describe the corpus design, annotation process, and we introduce preliminary experiments on sentiment analysis and topic modeling. The baselines can be used for new studies on this dataset and others. We do not focus on the current uses of Large Language Models for these tasks, but the corpus provided can still be useful for that approach in many ways: for instance it can be used as a part of a prompt or as an evaluation dataset.

With this work, we make the point that to have a single dataset with topic and sentiment information together is very helpful, since when it comes to sentiment analysis and opinion mining, it is essential to understand what is the subject of the stated opinion (Liu, 2012). We decided to work from the original B2W-Reviews-01 corpus for two main reasons: i. e-commerce reviewing is a textual genre in which popular, daily language is used, and it is driven to have explicit opinion and sentiments; ii. the initial work has much geographic and demographic information attached, such as gender, age and reviewer location, which can be useful for sociolinguistic analysis. This is not available in most of the machine-readable linguistic resources. Therefore, we believe that having a portion of the earlier B2W-Reviews01 dataset labeled for sentiment analysis and topic modeling can serve various purposes and be helpful to different perspectives. Although the present work can also be used to do aspect-based sentiment analysis (ABSA), we do not focus on this particular use, since the topics

presented in the data available are broader than the aspects of the product itself, as commonly targeted by ABSA (Zhang et al., 2022).

For those interested in e-commerce challenges, product reviews are an important source of information. It is essential to understand customers' negative and positive feelings in relation to their experience with a particular service or product. From the customers' perspective, the insights provided by reviews play a crucial role supporting others in their decision-making process (Zhang et al., 2023).

Due to the large volume of data generated by users every day, performing a manual analysis of this type of content is impractical. Thus, the use of automatic natural language processing techniques to analyze user-generated content (UGC) in a scalable and effective way has grown much in the last decade. Sentiment analysis and text categorization techniques have been widely used in the Brazilian industry, but there is a dearth of open labeled corpora in Portuguese containing data related to e-commerce.

We aim to improve this state of affairs, sharing with the NLP open-source/data community the RePro corpus. The corpus is freely available for non-commercial use on GitHub[1] and HuggingFace [2] under the license CC BY.NC.SA 4.0[3].

2 Related work

Following (Caseli and Nunes, 2023), we have, for Brazilian Portuguese, around ten different lexical resources for Sentiment Analysis and six available corpora. The OPCovidBR (Vargas et al., 2020) is the work most similar to ours. The corpus has 1,800 annotated tweets with topics (called "opinion groups") and polarity (positive or negative).

From now, we focus only on previous work both in Brazilian Portuguese and on review content since exploring the whole literature on topic modeling and sentiment analysis is not our focus.

The Brazilian Portuguese e-commerce genre was first described in the dataset 'Brazilian E-Commerce Public Dataset by Olist'[4]. `Olist` is a Brazilian marketplace which made available information about 100,000 orders between 2016 and 2018. This comprises real data, including order status, price, product attributes, and reviews written by customers.

The B2W-Reviews01 open corpus was introduced and made publicly accessible in 2019 through the efforts of (Real et al., 2019). The corpus B2W-Reviews01 is a publicly available collection of product reviews, comprising over 130,000 customer feedback entries sourced from the `Americanas.com` website during the period of January to May 2018. Notably, B2W-Reviews01 encompasses a wealth of information concerning the reviewers themselves, including aspects such as gender, age, and geographical location. Moreover, the dataset incorporates dual forms of review evaluation: the conventional 5-point rating scale, commonly represented by stars on e-commerce platforms, and a "recommend to a friend" label that requires a simple "yes" or "no" response, indicating the customer's inclination to endorse the product to others.

In 2020 a study was conducted (Real et al., 2020) to enhance the B2W-Reviews01 corpus by providing annotated samples, resulting in the creation of a new corpus named B2W-Reviews02. This supplementary corpus comprises 250 reviews extracted from the larger B2W-Reviews01 dataset. To gain comprehensive insights into customer opinions and sentiments expressed within these reviews, the authors approached the task as an aspect-based sentiment analysis (ABSA). This involved identifying the topics discussed within each review and analyzing the associated sentiment or polarity linked to each specific topic.

In the corpus brands.Br[5] (Fonseca et al., 2020), the authors conducted an annotation process for the same 250 samples that were annotated in (Real et al., 2020). Similarly, the authors considered an annotation of the topics approached in a review. Although the efforts

[1] https://github.com/lucasnil/repro
[2] https://huggingface.co/datasets/lucasnil/repro
[3] https://creativecommons.org/licenses/by-nc-sa/4.0/
[4] https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
[5] https://github.com/metalmorphy/Brands.Br

of (Real et al., 2020) and (Fonseca et al., 2020) may have insights about e-commerce annotation, the size of the annotated sample is not sufficient to train classical ML algorithms.

In 2021 the work described in (Zagatti et al., 2021) performed data anonymization procedures in the B2W-Reviews01 corpus to ensure compliance with the General Law for the Protection of Personal Data, the Brazilian legislation governing the processing of personal data.

UTLcorpus (Sousa et al., 2019) is a corpus with movies and apps reviews that also has 'helpful votes' information, users feedback about how helpful each review is. It has almost 3 million reviews and its main purpose is to tackle the lack of Helpfulness Prediction resources in Brazilian Portuguese.

3 Methodology

To create RePro, we randomly selected 10,000 reviews from B2WReviews-01, stratified by the star rate score. It means RePro has around 2,000 reviews of each point in the 1-5 star rating scale. We conducted a polarity and topic annotation, since user-generated content is not always reliable: it is common that the star rating score given by the user does not necessarily express the user sentiment described in the textual content.

We describe the annotation procedures and decision below.

3.1 Annotation guidelines

The elaboration of the annotation guidelines started with the exploration and analysis of the data, a stage when the most recurrent similar subjects are grouped into topics. This exploratory analysis helped to define six main groups concerning topic modeling: *advertising*, *product*, *delivery*, *receipt conditions*, *others*, and *inadequate*.

A summary of what each label represents is detailed below:

Advertising includes contexts in which the product delivered corresponds or not to the information displayed on the product page, for example, in the description, image, technical sheet, title, or to its advertising in general;

Product encompasses contexts related to quality, originality, cost-effectiveness, product attributes/characteristics, user experience, and also compliments in general;

Delivery refers to speed of delivery, delivery time, undelivered order, product pick-up at the physical store, virtual delivery (e.g. gift cards, code), and also remarks about shipping;

Receipt Conditions include contexts about the state of a product after the order is received, such as, for example, whether the product arrived damaged or not, well packaged (or not), defective products, incomplete orders, wrong/changed orders and assorted orders that meet (or do not meet) customer expectations;

Others are contexts related to questions to sellers, consumer service, stock, shopping experience, payment methods, meaningless information for other potential buyers but that are not harmful to the company;

Inadequate comprises harmful information, as profanity, mentions of competitors, legal references, external website links, personal information.

For sentiment classification, the polarity labels assigned were:

Positive, which characterizes reviews containing compliments or favorable comments in relation to products, services, or the company in general;

Negative, which characterizes reviews containing unfavorable comments or criticism;

Neutral, includes reviews without compliments and explicit criticism, such as questions regarding products, services, or the company in general;

Positive/Negative, which includes reviews containing both compliments and criticism in the same review.

A document discussing annotation guidelines, defining and detailing topics for each context, was prepared in order to serve as a guide for annotators. This is meant to minimize personal bias and ensure consistency and agreement in the data annotation phase. This document was tested, with a small batch of data, in order to level the understanding and measure the degree of agreement between the annotators before starting the official annota-

tion of the data.

3.2 Corpus annotation

The annotation task was multi-label for the classification of topics, that is, it admits more than one topic for the same review considering the topics described above. In cases of uncertainty between different topics, the orientation for annotators was to mark both topics in order to obtain a more general annotation. Regarding the sentiment annotation, the classification was multi-class, namely: positive, negative, neutral, and positive/negative. Given that a review is composed of a title and body, these two fields were taken into account for the annotation.

We had six annotators with previous e-commerce experience working in this process. All of them are Brazilian, from São Paulo, Minas Gerais, Rio de Janeiro and Santa Catarina states, and their first language is Brazilian Portuguese. Each sample was annotated by at least two annotators, a third specialist was responsible for curating and resolving all disagreements found in the initial annotation. The annotation batches were divided based on stars rating (1 to 5), expecting that, given the user score, each batch would have a stable nature which simplifies the annotation process.

At the end of each annotation round, we measured the Inter-Annotator Agreement(IAA) by Cohen Kappa coefficient (Cohen, 1960), and disagreements were sent to the curation. After curating each round, a meeting was held to provide feedback on disagreements, including new cases (non-existent in the data exploration sample), difficult cases (ambiguity, for example), and highlighting points of attention for the guideline criteria. On average, the IAA for topic annotation was found to be 0.68, while for sentiment annotation, the average Cohen's Kappa reached a value of 0.71. The present values serve as indicators of the extent of concordance observed among human annotators, with elevated Cohen's Kappa coefficients suggesting heightened levels of agreement. In our investigation, the achieved values signify a substantial level of agreement for both topic and polarity annotations, demonstrating the trustworthiness and uniformity of the annotation process across numerous iterative cycles.

3.3 Results

Here we present general information of RePro. This corpus contains 10,000 samples, labeled with 6 different topics, each sample may have one up to six topics. Figure 1 shows the distribution of samples by topic.

Considering the polarity/sentiment annotation, we have 4 possible labels, each sample is labeled with only one of them. Figure 2 shows the distribution of samples by sentiment polarity.

The corpus is released in `csv` format with all the previous information available in B2W-Reviews01. Thus it has the following columns: **A**: submission_date; **B**: reviewer_id; **C**: product_id; **D**: product_name; **E**: product_brand; **F**: site_category_lv1; **G**: site_category_lv2; **H**: review_title; **I**: review_text; **J**: overall_rating; **K**: recommend_to_a_friend; **L**: reviewer_birth_year; **M**: reviewer_gender; **N**: reviewer_state; **O**: topics (a list of all the topics found in this review); **P**: polarity.

There are no `null` values for `topics` and `polarity` columns.

To facilitate data analysis, the topics listed in column **O** are further distributed across columns **Q** to **V** in the specified order: *delivery*, *others*, *product*, *receipt conditions*, *inadequate*, and *advertising*. These columns can be assigned a value of 0 or 1, indicating the absence or presence of the respective topic.

To make it clearer, the following is an example of a sample:

```
A: 2018-01-11 08:33:53
B:cb0468b5ce0aa0a2f5 (etc...)
C:132743826
D: Jogo de Cama Casal Liz 4
Peças - Corttex
E:
F: "Cama, Mesa e Banho"
G: Jogo de Cama
H: ..
I: Gostei muito o preço esta
bem em conta Eu recomendo.
J: 3
```

```
K: Yes
L: 1997
M: F
N: MG
O:['PRODUTO']
P:['POSITIVO']
Q: 0
R: 0
S: 1
T: 0
U: 0
V: 0
```

The `reviewer_id`, column B, is longer than we can display here. Column E, `product_brand` is empty, since the brand of the product was not available in the initial corpus, but for those interested in brands, it is often possible to infer the product brand from the product title. In this review, the reviews just leave `..` as a `review_title` (column H). The `review_text` in column I contains the detailed text of the review, with this example expressing satisfaction with the product's pricing: "Gostei muito o preço esta bem em conta Eu recomendo"[6]. The text in RePRO is exactly the text written by the reviewers, without any treatment. Column M, `reviewer_gender` has possible values among M (masculine) and F (Feminine), and few instances are empty[7]. In column N, we find the acronyms for the Brazilian States, this column can be empty. Column O, `topics`, presents a list of topics associated with the review; here, it is ['PRODUTO'] (product). Column P, `polarity`, indicates the sentiment polarity of the review, labeled as ['POSITIVO'] (positive). Finally, columns Q to V correspond to the distribution of specific topics across these columns, where a value of 1 or 0 signifies the presence or absence of the respective topic. In this example, "Product" (Q) is marked with a value of 1, while others remain at 0.

[6]"I liked very much, the price is well worth it, I recommend."

[7]Note that the corpus was collect in 2018, when the gender discussion were not as vivid as today. Today it is possible to not inform the user gender in the registration in Americanas.com. However, there are still only these two possible gender options in the registration form.

Figure 3 depicts the distribution of sentiment polarity categories (positive/negative, positive, negative, and neutral) across different overall ratings (1 to 5). The distribution of overall ratings varies among the sentiment categories. For "positive/negative" sentiment, ratings are predominantly clustered around the middle range, with the highest concentration of reviews rated 3 (751 reviews) and 2 (621 reviews). Still, some reviews in this polarity received the highest rating of 5 (154 reviews). This suggests that customers expressing mixed sentiments are more inclined to provide average ratings rather than extreme ones. Similarly, reviews with "neutral" sentiment also tend to receive ratings mainly in the mid-range, with 239 reviews rated 3, 96 reviews rated 2, 44 reviews rated 1, and 17 reviews rated 4. There are only 13 reviews with "neutral" sentiment receiving the highest rating of 5, suggesting that neutrality also tends to correlate with mid-range ratings. In contrast, for purely "positive" sentiment, ratings are more dispersed, with a large number of reviews receiving high ratings of 4 (1,598 reviews) and 5 (1,819 reviews), indicating that customers expressing positive sentiment are more likely to award higher ratings. However, it is noteworthy that there is a minimal number of reviews labeled as "positive" which received the lowest rating (4 reviews). For "negative" sentiment, the distribution is skewed toward the lowest ratings, with a significant number of reviews rated 1 (1,825 reviews) and 2 (1,257 reviews). However, in contrast, very few reviews in this category received higher ratings, indicating a general connection between negative sentiment and low overall ratings. It's worth noting that while this connection is prevalent, there are hard cases, especially considering the three star rate, in which around 17% of the users feedback are labeled as positive. It suggests that ratings may not always be entirely reliable, particularly when considering the use of ratings as labels for training a supervised machine learning model in the task of sentiment polarity classification.

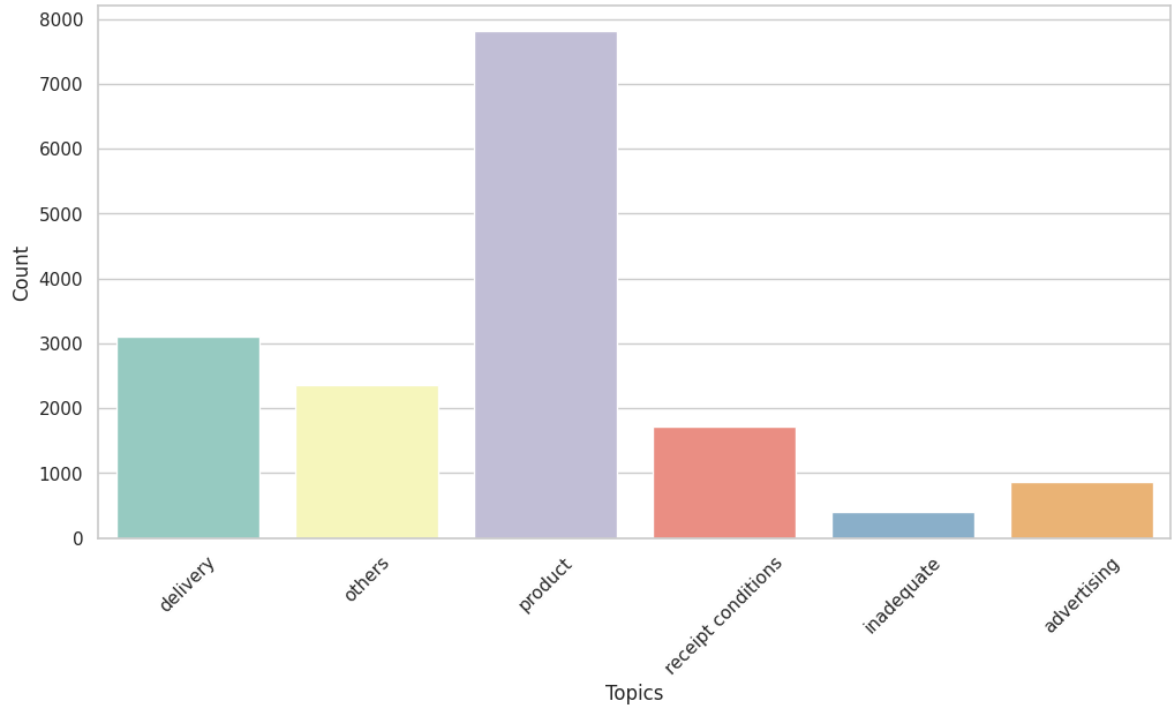

Figure 1: Distribution of samples by topic

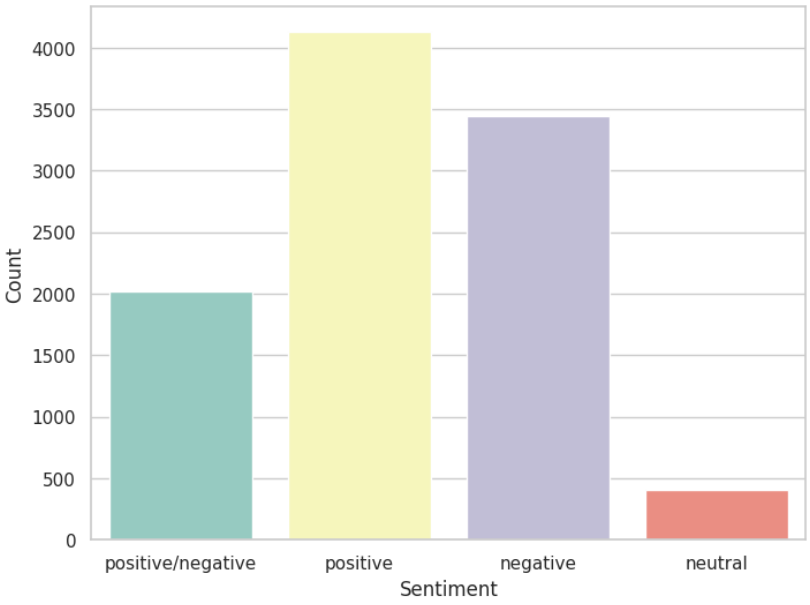

Figure 2: Distribution of samples by sentiment polarity

4 Corpus Evaluation

In this section, we outline a simple experiment aimed at assessing a machine learning model's proficiency in executing the designated tasks within RePro[8]. To accomplish this, we utilized a cutting-edge model built upon the Transformer architecture. Specifically, we employed a pre-trained Portuguese-language BERT model known as BERTimbau(Souza et al., 2020).

Initially, we performed a random split of the dataset into training and test sets. In this regard, 70% of the data was allocated for fine-tuning the model, while the remaining 30%

[8]The code to reproduce this experiment is available on: https://github.com/lucasnil/repro

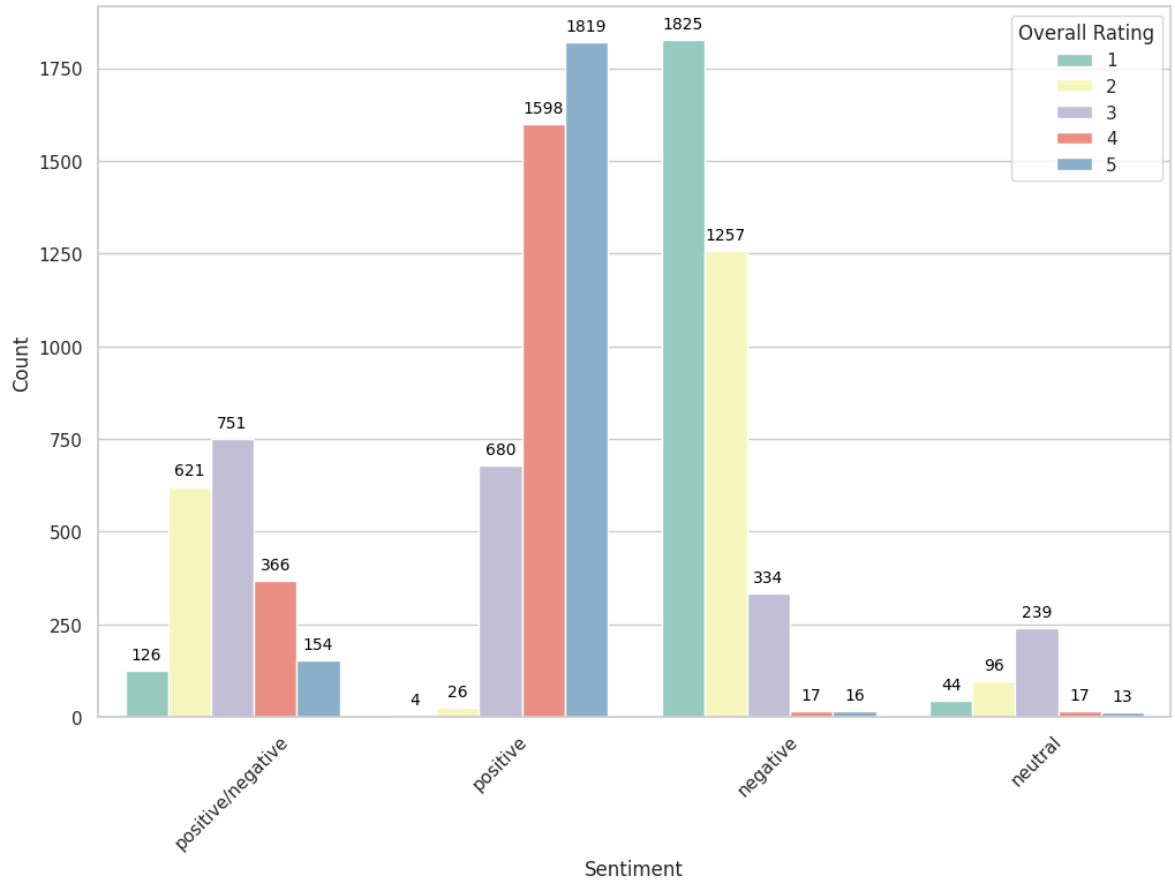

Figure 3: Distribution of sentiment polarity by overall rating

Classes	Precision	Recall	F1-score	Samples
Advertisement	0.92	0.90	0.91	273
Delivery	0.96	0.99	0.97	887
Product	0.96	0.98	0.97	2347
Receiv. cond.	0.92	0.88	0.90	501
Inadequate	0.77	0.52	0.62	121
Others	0.88	0.89	0.89	723
Average	0.90	0.86	0.88	

Table 1: Results obtained from the topic categorization task on the test dataset.

Classes	Precision	Recall	F1-score	Samples
Neg./Pos.	0.89	0.86	0.88	598
Negative	0.94	0.95	0.95	1056
Neutral	0.88	0.81	0.84	129
Positive	0.96	0.97	0.96	1218
Average	0.92	0.90	0.91	

Table 2: Results obtained from the sentiment classification task on the test dataset.

was reserved for evaluating its generalization ability on unseen samples. We fine-tuned one model for each task. For both models, we used AdamW (Loshchilov and Hutter, 2017) as the optimizer, with a learning rate of 4e-5 and a batch size of 8. The topic categorization model underwent ten epochs during training, while the sentiment classification model was trained for seven epochs.

The results obtained for the topic categorization and sentiment classification tasks are presented in Tables 1 and 2, respectively.

The findings regarding the SA task were promising, as indicated by F1 Scores equal to or exceeding 0.84. However, it is noteworthy that the model demonstrated relatively lower performance in distinguishing between the `[POSITIVE, NEGATIVE]` and `[NEUTRAL]` classes. This discrepancy may be attributed to the inherent ambiguity associated with characterizing these classes in comparison to the relatively more distinguishable `[POSITIVE]` and `[NEGATIVE]` classes.

In the task of TM, the obtained results were generally satisfactory, except for the `[INADEQUATE]` class. The F1 scores for most classes were 0.89 or higher, indicating

that the model successfully learned to classify these topics accurately. However, the [INADEQUATE] class exhibited lower performance, which we discuss below.

4.1 Error Analysis

To perform an error analysis, we manually reviewed and categorize 100 randomly selected samples for TM task and 50 samples for SA. At the end of the error analysis, we manually investigated and categorized 150 samples, making sure that we reviewed all the possible combinations of misclassification. We analyzed more samples of TM since there are more label combinations for this task.

Considering the SA task, from 50 samples, the most common error was related to the presence of adversative coordinating conjunctions used in contexts in which the opposition was not related to the quality of the product/service, but used to emphasize a specific aspect or topic of the main text. We counted 11 errors, so more than 20% of the analyzed mistakes were related to it. One example of it is the following sample: "Bom custo benefício. Não surpreende, **mas** vale muito o valor pago por ele. Não sou especialista, **m**as acho ótima a resolução e a sensibilidade da tela."[9], which was annotated as [POSITIVE] and predicted as [POSITIVE, NEGATIVE].

For TM, for 40% of the errors, the model successfully predicted some of the expected classes but not all of them. Unsurprisingly, the model struggles to correctly categorize the topics [OTHERS] and [INADEQUATE]. To illustrate, in "Gostei. Gostei do produto, tive um problema com assistencia mas foi rapidamente resolvido"[10], annotated as [PRODUCT,OTHERS], the model could only correctly predict the class [PRODUCT].

It is important to highlight that the class [INADEQUATE] is the one with less examples in the corpus, while the class [OTHERS] comprises different sub-topics. Also, both of them frequently co-occur with other categories, so it was expected that their classification would prove to be particularly challenging.

[9] Good cost-benefit. It's not surprising, **but** it's well worth the price paid for it. I'm no expert, **but** I think the resolution and sensitivity of the screen are great.

[10] "I liked it. I liked the product, had an issue with customer support, but it was quickly resolved."

5 Conclusion

In this work, we described RePro, a 10,000 samples of e-commerce product reviews in Brazilian Portuguese, manually annotated with polarity and topics. We aimed to have a detailed description of the annotation process, since this corpus can be used as a benchmark for future work.

We also provided preliminary experiments for topic modeling and sentiment analysis based on BERTimbau, a pre-trained Portuguese-language BERT system. Our goal was not to exhaustively test different algorithms and architectures for these two tasks, but rather to provide reproducible baselines for future work.

With this work, we target to improve the Natural Language Processing scenario for the scholars' community, that still struggles to find high quality open data to investigate Portuguese processing.

References

H. M. Caseli and M. G. V. Nunes, editors. 2023. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN. https://brasileiraspln.com/livro-pln.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

B. Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan &. Claypool Publishers.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Livy Real, A Bento, K Soares, Marcio Oshiro, and Alexandre Mafra. 2020. B2w-reviews02, an annotated review sample. In *OpenCor*.

Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01-an open product reviews corpus. In

the Proceedings of the XII Symposium in Information and Human Language Technology, pages 200–208.

Rogério Figueredo de Sousa, Henrico Bertini Brum, and Maria das Graças Volpe Nunes. 2019. A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Symposium in Information and Human Language Technology - STIL*. SBC.

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)*.

Francielle Alves Vargas, Rodolfo Sanches Saraiva Dos Santos, and Pedro Regattieri Rocha. 2020. Identifying fine-grained opinion and classifying polarity on coronavirus pandemic. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 511–520. Springer.

Fernando Zagatti, Lucas Silva, and Livy Real. 2021. Anonymization of the b2w-reviews01 corpus. In *OpenCor*.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2023. Dive into deep learning.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.