InferBR: a Natural Language Inference Dataset in Portuguese

Luciana Bencke, Francielle V. Pereira, Moniele K. Santos, Viviane P. Moreira
Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre - Brasil
{lrbencke, fvpereira, mksantos, viviane}@inf.ufrgs.br

Abstract

Natural Language Inference semantic concepts are central to all aspects of natural language meaning. Portuguese has few NLI-annotated datasets created through automatic translation followed by manual checking. The manual creation of NLI datasets is complex and requires many efforts that are sometimes unavailable. Thus, investments to produce good quality synthetic instances that could be used to train machine learning models for NLI are welcome. This work produced InferBR, an NLI dataset for Portuguese. We relied on a semiautomatic process to generate premises and an automatic process to generate hypotheses. The dataset was manually revised, showing that 97.4% of the sentence pairs had good quality, and nearly 100% of the instances had the correct label assigned. The model trained with InferBR is better at recognizing entailment classes in the other Portuguese datasets than the reverse. Because of its diversity and many unique sentences, InferBR can potentially be further augmented. In addition to the dataset, a key contribution is our proposed generation processes for premises and hypotheses that can easily be adapted to other languages and tasks.

Keywords: natural language inference, synthetic data, text generation

1. Introduction

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is a classification task focused on deduction (de Souza Salvatore et al., 2023) – a model is presented with a pair of sentences and classifies the relationship between their meanings (Jurafsky and Martin, 2023). The first sentence is known as the *premise* (P), and the second is the *hypothesis* (H). An NLI model should infer whether (i) H entails P (*i.e.,* based on P, we can infer H is true), (ii) H contradicts P (*i.e.,* based on P, we can infer H is false), or (iii) H is neutral in relation to P (the truth of H cannot be determined on the basis of P).

Understanding these relationships is fundamental because their semantic concepts are central to all aspects of natural language meaning (Bowman et al., 2015a; Van Benthem, 2008; Katz, 1972). Learning to classify how H relates to P is useful for the development of semantic representations (Bowman et al., 2015a). Additionally, since comprehending a piece of text means knowing whether it is true, training good models for NLI is key for semantic systems (Marelli et al., 2014).

When considering the volume of digital texts available, Portuguese is not a low-resource language, but there are two NLI annotated datasets: SICK-BR (Real et al., 2018) and ASSIN2 (Real et al., 2020). Data augmentation strategies can be used to generate synthetic instances for classification tasks (Bencke and Moreira, 2023; Yoo et al., 2021; Bayer et al., 2023). With recent astonishing advances in Large Language Models (LLM) highly specialized in following human instructions, text generation established itself as a viable option to generate synthetic data for several tasks. However, the difficulty in generating synthetic instances for NLI is not only related to *how* to generate such pairs but also if they would be correctly labeled respecting the entailment, contradiction, and neutral concept boundaries. The challenge is to explore semantic variations and still maintain the entailment classes. According to Sadat and Caragea (2022), creating new NLI datasets capturing linguistic properties of different domains is complex and sometimes impossible. Thus, efforts towards reducing the reliance on manually annotated data in training deep learning models for NLI are welcome.

Aiming to contribute to language resources for Portuguese, this paper introduces InferBR, a Portuguese NLI dataset semi-automatically generated using the Generative Pre-trained Transformer GPT4 but with humans revising the synthetic instances. The generation process was conducted separately, first to generate the premises and secondly the hypotheses. The premises are created by transforming some training instances from existing resources in Portuguese: i) a recent image-to-text dataset in Portuguese named PraCegoVer (in English, would be "ForTheBlindToSee") (dos Santos et al., 2022) and ii) a small sample of an existing NLI dataset was used as source to generate very diverse premises through GPT-4. For both groups of premises, the hypotheses were generated through a prompt engineering strategy. For evaluation, two reviewers analyzed the instances. They judged whether the premise and hypothesis were coherent and if the label was cor-

LREC-COLING 2024, pages 9050–9060
20-25 May, 2024.

rect, with a third reviewer (R3) voting in cases flagged as dubious. The general conclusion is that 97.4% of the generated pairs had good quality; among those, 99.9% had the correct label. We also ran inter-dataset experiments crossing models and test sets with other NLI datasets. The results showed the model trained in InferBR was better at recognizing entailment classes in the other datasets than the reverse. This may indicate greater generalization capabilities.

Besides generating a language resource, this paper contributes to the presentation of processes for generating such datasets, which may also be adapted for tasks other than NLI and be reused by researchers working on low-resource languages. We made the dataset available[1], indicating instances flagged as low quality after human validation and the generated labels.

2. Related work

We investigated two main groups of works: the prominent NLI datasets in English and Portuguese and recent works producing synthetic NLI instances either automatically or semi-automatically. In special, we considered works that use the text-generation capabilities of LLMs.

The Stanford Natural Language Inference (SNLI) (Bowman et al., 2015b) dataset is a collection of 570k pairs of sentences labeled for entailment, contradiction, and semantic independence. Premises were collected from Flickr30k, a dataset with image captions (Young et al., 2014a). The authors used Amazon Mechanical Turk to collect the hypotheses, and about 2,500 workers contributed to the task. The premises were presented to the workers, who were asked to write hypotheses for each label (entailment, neutral, and contradiction). Four annotators revised ten percent of the pair labels, achieving agreement in almost 98% of the revised sample.

Another important English dataset is the Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018), which contains 433k pairs of sentences. Premises were derived from ten sources representing ten different genres (government, letters, fiction, etc.). Generating the hypotheses was similar to SNLI, and four annotators confirmed each label. The authors argued that this diversification captures more of the complexity of modern English.

The English dataset SICK (Marelli et al., 2014) (Sentences Involving Compositional Knowledge) with 10k instances served as the basis for the two publicly available NLI datasets in Portuguese. SICK was built by sampling sentence pairs from 8K ImageFlickr and SemEval 2012 STS MSRVideo description dataset. First, sentences were normalized to remove undesirable linguistic phenomena; then, they were expanded using syntactic and lexical transformations to obtain up to three new sentences with specific features. Finally, all the sentences generated in the expansion phase were paired with the normalized sentences to obtain the final dataset. Humans annotated the relatedness in meaning and entailment of each sentence pair. This annotation process was done through a sizeable crowdsourcing initiative using platforms such as Amazon Mechanical Turk, where ten different humans evaluated each pair, and the order of presentation of the sentences was counterbalanced.

Portuguese NLI datasets are usually translations. The dataset SICK-BR (Real et al., 2018) is the manually revised translation of the English SICK(Marelli et al., 2014).

Ten annotators thoroughly revised all translations. An online tool was used in the annotation process and annotators looked at each other's work when translating their sentences. The authors' main goal was to keep pairs of sentences from English and Portuguese aligned as much as possible, preserving the labels. We use part of SICK-BR as a source to generate synthetic data; more details about it are described in Section 3.1.

ASSIN2 (Real et al., 2020) is the Portuguese acronym for Evaluating Semantic Similarity and Textual Entailment and it corresponds to an NLI dataset with two classes: entailment and non-entailment. It used SICK-BR data and, aiming at balanced classes, authors used a semi-automated strategy, taking SICK-BR pairs annotated as entailment and changing some synonyms or removing adverbial or adjectival phrases. They also created new pairs for the entailment class. All generated pairs were annotated by at least four native speakers of Brazilian Portuguese with linguistic training. Only pairs annotated with the same label by most annotators were included in the dataset.

Considering the works that generated synthetic instances, Akoju et al. (2023) used data augmentation techniques to produce a new synthetic dataset with 1,304 sentence pairs created by modifying 15 examples from the SICK English dataset. They used a variety of modifiers (universal quantifiers, existential quantifiers, negation, etc.). Results were evaluated using NLI models trained in other datasets to predict the instances and fine-tune the same models on the new dataset. They did not find significant differences in the results, but analyzing the predictions, they observed that instances modified with adjectives, adverbs, and universal quantifiers performed better than sentences modified with negation and existential quantifiers. Additionally, the authors report that

[1] https://github.com/lbencke/InferBR

models seem confused when the label is Neutral, and the modifier types are negations.

Text generation was explored by Sadat and Caragea (2022) to create hypotheses for selected premises obtained from other NLI datasets in English. The authors proposed a semi-supervised learning (SSL) framework. First, they fine-tuned BART models with a small set of pairs, conditioning them to produce hypotheses for each class: the premises from the selected pairs are used as the source texts, and their hypotheses as the targets. Ultimately, they got one conditioned BART^C model per class C. With BART^C they generated hypotheses for given premises and assigned the pseudo-labels of each class C. They selected only instances with high-confidence predictions and added them to the labeled dataset, repeating the process until a defined limit. They compared results on the produced instances with BERT models trained with fully human-annotated datasets, with some data augmentation methods and other SSL methods getting superior performance.

We can conclude from the investigated works in English and Portuguese that they involved a vast manual effort to produce important datasets. The work by Sadat and Caragea (2022) is close to ours since it approaches the creation of NLI datasets automatically using text generation. Our work differs from theirs because we have not fine-tuned one language model to be conditioned to each class. We took advantage of GPT-4's ability to follow instructions due to its training on reinforcement learning from human feedback (Christiano et al., 2017), and using an in-context learning strategy (Brown et al., 2020) designing prompts with qualitative selected examples at inference time. Regarding evaluation, we also differ from them since we manually revised all the generated pairs, which gave us a real perception of the quality of our dataset.

3. Dataset generation

Text generation refers to producing sequences of text conditioned on an input text (prompt) (Liu et al., 2021; Zhang et al., 2022). Deep learning autoregressive language models are usually applied in these tasks. These models are trained on Causal Language Modeling (Zhu et al., 2023), where each produced token is added to the sequence of inputs, generating the next token based on that input, and so on;

Currently, there has been an avalanche of LLMs. We chose the Generative Pre-trained Language models family from OpenAI since we wanted to deliver our first effort evaluating a paid resource output as a possible raw material for developing NLI datasets. The research path from GPT(Radford et al., 2018), GPT-2(Radford et al., 2019), and GPT-3(Brown et al., 2020) and GPT-4 (Achiam et al., 2023) leverages more data and computation, which, as advocated by OpenAI, is necessary for the development of increasingly sophisticated and capable language models. The first two models were open, but since GPT3, the company made the model available through an API.

We used two datasets as sources to produce the premises and to design prompts submitted to GPT-4 to generate the three types of NLI hypotheses. The generated dataset was manually revised. In all generative tasks, we set the temperature to 0.8 for decoding. The cost of using GPT-4 via the Open API was USD 220. This amount also includes trials that were not used in the final dataset. The next sections describe in detail the processes to generate premises and hypotheses.

3.1. Generating Premises

We want to generate premises that are comprehensible, unambiguous (any person would understand the same from it), informative (do not need additional data to understand it), coherent (logically structured, easy to read), and that respect common sense. We used two datasets as sources for the premises: PraCegoVer (dos Santos et al., 2022) and SICK-BR (Real et al., 2018). The process used to get the final premises from each dataset was different, as shown in Figure 1.

3.1.1. PraCegoVer

PraCegoVer (dos Santos et al., 2022) is a multimodal dataset with images and their descriptions in Portuguese. It came from a social movement that started in 2012, aiming to increase the inclusion of people with visual impairments, and its goal was to contribute to developing models to automate image captioning. The initiative stimulated social Instagram users to post images tagged with #PraCegoVer and add a short description of their content. The image captions have 39.3 words on average, and the standard deviation is 29.7, this is considered very challenging for image captioning tasks compared to other datasets such as Flickr30k (Young et al., 2014b) and MS-COCO (Lin et al., 2014), which have, on average, 10 words per caption and low variance.

This dataset contains many named entities, including several brand names. This is because companies used the #PraCegoVer to post messages describing images of their products. Even people's names are in the dataset because users would post photos with friends and include their names in the descriptions.

The process for generating the premises using PraCegoVer is depicted in Figure 1a. To filter out descriptions with proper nouns, in Step (1), we

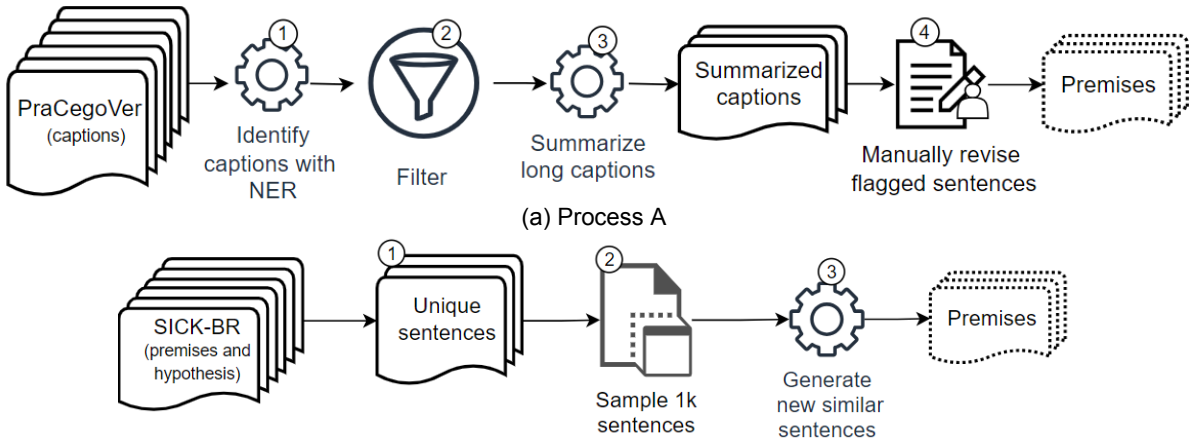

Figure 1: Premise generation process.

run a model trained for Named Entity Recognition (NER)[2]. Instances containing entity names are flagged and then removed in Step (2), where we also remove special characters. In Step (3), we used GPT4 to summarize the filtered image captions since they contain long descriptions with several details, and we wanted to reduce them to obtain a concise premise. The instruction sent to the model requested a summary of the scene, avoiding mentioning it is about a picture (since several descriptions have those mentions when describing the image). Some examples of the original caption (O) and the generated summary (G) can be seen in Table 1. It is worth pointing out that some captions in PraCegoVer have grammar and orthography errors – that is the case of example 01 in which there is a missing verb between "*possível*" and "*os*". These issues end up being corrected by the summarization step. Finally, in Step (4), we revise premises flagged with the following issues: errors pointed out in a spell-checker or containing some expressions that could make the premise confusing and that are repeated in several descriptions like *"retratado"* ("*portrayed*", in English), etc. During this phase, around 25% of the generated premises were revised because they were flagged for some of the mentioned issues, and almost 60% of this group was modified to yield coherent text.

01) **O**: As Fotos mostram o espaço do bicicletário por diversos ângulos. É possível os espaços onde as bikes devem ser alocadas. *The photos show the bike rack from different angles. Spaces where bikes should be allocated are possible.* **G**: O bicicletário é exibido de vários ângulos, destacando os locais designados para a alocação das bicicletas *The bike rack is displayed from multiple angles, highlighting designated bike storage locations*
02) **O**: Foto em preto e branco com imagem do apresentador no palco, com microfone na mão. *Black and white photo with image of the presenter on stage, with microphone in hand.* **G**: O apresentador está no palco segurando um microfone *The presenter is on stage holding a microphone*
03) **O**: foto de porquinhos recém nascidos, de olhos fechados e a frase "Somos só bebês" *photo of newborn piglets, with their eyes closed and the phrase "We are just babies"* **G**: Recém-nascidos porquinhos de olhos fechados são identificados como apenas bebês *Newborn piglets with their eyes closed are identified as just babies*

Table 1: Examples of the summarization step. (O) corresponds to the original image caption; (G) is the generated summary. We also show the translation into English in italics.

3.1.2. SICK-BR

SICK-BR (Real et al., 2018) originated from the translation of the English dataset SICK (Marelli et al., 2014). Figure 1b presents the process of generating premises from SICK-BR. In Step (1), we extracted the unique sentences from the training and validation splits. There are 4.5k unique premises and 4.6k unique hypotheses. 3.2k sentences appear both as a premise and hypothesis, but they are paired with different combinations to make up unique pairs. We created a single set with 6k unique sentences using training and validation sets of SICK-BR. From this final set, in step (2), we sampled 1k sentences to be used as sources to generate others. In Step (3), we prompted GPT4 to write new instances with similar

[2] https://huggingface.co/monilouise/ner_news_portuguese

contexts; this instruction ensures semantic variations that allow for greater diversity than just paraphrasing. Table 2 presents some examples of the source and the generated instances.

O: Muitas crianças estão de pé *Many children are standing* **G**: Vários jovens estão sentados *Several young people are sitting*
O: Não tem nenhum cachorro perseguindo uma bola *There is no dog chasing a ball* **G**: Não há nenhum gato brincando com o novelo de lã *There is no cat playing with the ball of yarn*
O: Uma pessoa está imprudentemente montando um cavalo *A person is recklessly riding a horse* **G**: Um indivíduo está temerariamente conduzindo um automóvel *An individual is recklessly driving a car*
O: Quatro pessoas estão paradas silenciosamente ao ar livre *Four people are standing silently outdoors* **G**: Três indivíduos permanecem quietos sob a luz da lua *Three individuals remain quiet under the moonlight*

Table 2: Premises generated from sample sentences in SICK-BR. (O) corresponds to the original sentence, and (G) is the sentence generated by GPT4. In italics, the translation to English.

3.2. Generating Hypotheses

We used a few-shot inference strategy, depicted in Figure 2. We first build a list of 50 examples to be used in the prompt, each corresponding to the layout indicated in number (1) of the flow. In step (2), for each premise, we sample three examples from the 50 available and add them to the prompt. The examples used for each premise may vary, which is desirable to add diversity to the dataset. Step (3) presents the structure of the prompt, which has four components: (i) an initial request – the same for all premises, (ii) short descriptions for the three classes, (iii) the three selected examples, and (iv) the premise for which the hypotheses should be generated. In Step (4), we used GPT4 to generate the hypotheses. The output is three labeled hypotheses that, along with the premise, compose the three labeled pairs for the dataset.

4. The InferBR dataset

InferBR was created using two strategies to generate the premises: 41% of the data was generated using process B with SICK-BR premises as the source to create new instances with a similar context but more aggressive than paraphrasing to yield more diversity. The majority of the data, 59%, came from PraCegoVer using Process A. The rationale was to have more instances that were not present in the other two NLI datasets. We generated the sets for training and testing, and Table 3 presents the numbers per split and class for InferBR and the other two existing NLI datasets in Portuguese. In our dataset, the premises in training and validation sets do not appear as premises in the test set.

	train	val	test
InferBR			
Contradiction	2,800	215	586
Entailment	2,799	216	586
Neutral	2,800	215	586
10803			
SICK-BR			
Contradiction	998	224	202
Entailment	1,948	437	436
Neutral	3,941	815	839
Total: 9840			
ASSIN2			
Non-entailment	3,250	250	1,224
Entailment	3,250	250	1,224
Total: 9448			

Table 3: Number of instances per split and class for the NLI datasets in Portuguese

The premises in InferBR were longer compared to SICK-BR, as can be seen in Figure 3, and the generated hypotheses were shorter than the premises. This reduction of the hypotheses also occurs in the English benchmark SNLI (Bowman et al., 2015a).

We checked for duplication of premises and hypotheses both within the individual datasets and across datasets. The results are in Table 4. InferBR has fewer unique premises (P) but many more than twice the number of unique hypotheses (H). In addition, there is almost no duplication when analyzing unique sentences in the dataset ($P \cup H$). We see this as an advantage since we ensure more diversity to the hypotheses and also may allow us to apply future augmentation techniques to combine similar H from this larger set in different ways, improving the learning process. It is worth pointing out that, although InferBR used SICK-BR as a source for premises and hypotheses, the resulting dataset has a very small overlap with SICK-BR (8 and 4, respectively). This happens because our process is able to change the original context.

Table 4 shows that, as described in the original paper, most of the sentences in ASSIN2 come

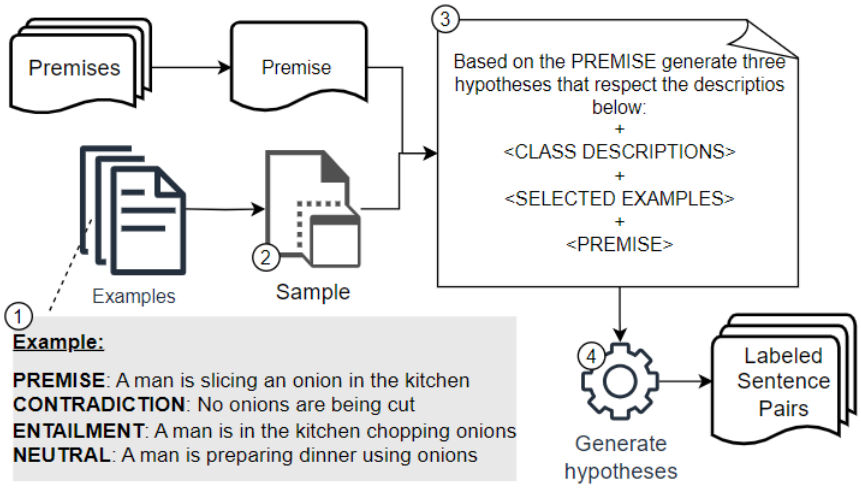

Figure 2: Hipotheses generation process.

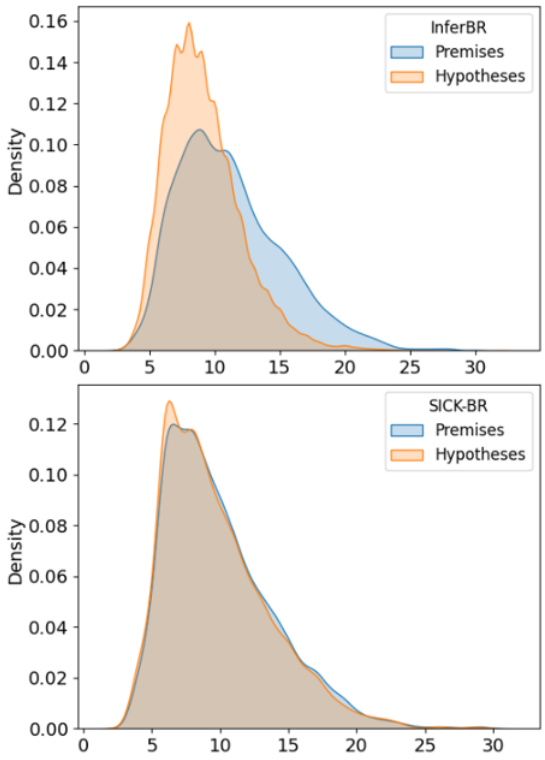

Figure 3: Distribution of tokens

Set	P	H	$P \cap H$	$P \cup H$
ASSIN2	5,150	5,172	3,814	6,508
SICK-BR	5,001	4,929	3,846	6,084
InferBR	3,600	10,669	450	13,819
ASSIN2 ∩ SICK-BR	3,901	3,794	2,642	5,053
InferBR ∩ SICK-BR	8	4	1	11
InferBR ∩ ASSIN2	10	4	2	12

Table 4: Unique premises (P), hypotheses (H), sentences ($P \cup H$), and their intersection ($P \cap H$) calculated intra- and inter datasets.

from SICK-BR. The pairs in ASSIN2 were reorganized to ensure the balance between the two classes (entailment and non-entailment), different from SICK-BR, which has three classes but is imbalanced.

To analyze the Part-Of-Speech (POS) categories in the datasets, we ran a POS tagger from *spaCy*[3] using the largest model available `pt_core_news_lg`. For each category, we calculated the average occurrence in the premises and hypotheses in InferBR and SICK-BR.

The results are presented in Table 5, where we can see that InferBR has more adjectives in the premises. This is expected because PraCegoVer instances refer to many brands' advertisements and people expressing opinions on the social network when describing the image ("*beautiful t-shirt*", "*gentle body lotion*"). We also see more nouns in InferBR, which is naturally related to the longer sentences, but also because it presents more detailed descriptions, for example, dishes ("*shrimp with tomatoes and grilled okra*") or what people are wearing ("*The person is getting ready to go out in her red bodysuit, striped pants, and red sandals*"). Both premises and hypotheses have a lower incidence of auxiliary verbs in InferBR. The occurrence of punctuation is higher in InferBR because it maintains the period in many instances, and more cases have a comma: "In the drinks cabinet, there are three bottles with labels.". InferBR's hypotheses have fewer determiners. This is associated with more cases where the article was replaced by the number ("*two girls are smiling*"), to a higher number of existential constructions (Mc-

[3]https://spacy.io

Nally, 2021) ("*There are people at the geological site*."), and also to more presence of mass nouns ("*Water is coming out of the beach shower*").

POS	Premise		Hypothesis	
	InferBR	**SICK-BR**	**InferBR**	**SICK-BR**
adjective	0.79	0.60	0.58	0.60
adposition	1.80	1.46	1.28	1.45
adverb	0.23	0.32	0.24	0.30
auxiliary	0.85	1.08	0.78	1.08
coord.conj.	0.24	0.23	0.11	0.24
determiner	1.84	1.84	1.54	1.84
noun	3.43	2.97	2.73	2.96
numeral	0.13	0.12	0.05	0.12
pronoun	0.14	0.11	0.12	0.11
proper noun	0.09	0.06	0.06	0.06
punctuation	0.23	0.05	0.52	0.05
subord.conj.	0.08	0.01	0.06	0.01
verb	1.31	1.24	1.09	1.23

Table 5: Average occurrence of each POS class in premises and hypotheses.

Cohen's Kappa
Between R1 and R2: 0.9693
Between R1 and G: 0.9820
Between R2 and G: 0.9842
Validation Statistics
Number of pairs validated: 10,803
Agreements: 10,538
Dubious cases checked by R3: 93
Flagged as low quality: 275
Agreements (errors in generated label): 10
Disagreements: 184
Confusing text: 81
Premises from Process A: 51
Premises from Process B: 15
Hypotheses: 15

Table 6: Statistics on the manual validation by reviewers (R) comparing results to the generated text and label (G).

5. Results

The evaluation strategy involves manually revising the dataset and comparing the performance of models trained with the generated data in recognizing entailment using the other two datasets in Portuguese and vice versa.

5.1. Human Validation

Two MSc students with ongoing research in the NLP field evaluated all the pairs and labels. They were unaware of the processes that generated the premises and hypotheses, and they also did not access each other's annotations while the revision was ongoing. They annotated the label they understood as correct for each pair and judged if the text of premises and hypotheses was comprehensible, clear, unambiguous, coherent, and according to common sense. If the text did not have all these characteristics, they flagged it as "confusing". When reviewers were unsure about the label, they were oriented to flag the instance. Whenever an instance was flagged by at least one reviewer, it was checked by a third reviewer (R3) who also assigned labels to these instances.

Table 6 summarizes the manual validation results. There is a very high agreement between reviewers and an even higher agreement between each reviewer's annotated label and the automatically generated one. Among the instances in which reviewers agreed and did not find any issue with the text, most of the labels assigned during the hypothesis generation process are correct (99.9%). The ten errors found are related to neutral boundaries with entailment and contradictions.

Only 2.6% of the instances were flagged as low-quality: 70% of those were hypotheses where the reviewers disagreed on the label, and no instances were flagged for R3 to check. Most of these instances are related to unclear boundaries between *Neutral* and the other two classes. 30% of the low-quality pairs are premises or hypotheses that were confusing for at least one reviewer and R3.

In total, only 0.75% of the pairs contain confusing text. It happens more in the premises because a problem in the premise affects three instances with hypotheses derived from it. The root cause of the problem can be in processes A or B. Confusing text generated by process A is mainly related to describing unreal images or specific parts of a bigger scene (Example 02 in Table 7). On the other hand, the confusing premises generated by Process B are mostly cases where GPT-4 faced issues with common sense or world knowledge (Example 01: parrots usually imitate the human voice, and one could picture it in a microphone, but the cat meowing on the megaphone is not a good similar context choice). The confusing hypotheses are also related to GPT-4. Besides common sense issues, there are some problems with the fluidity of the text (in Example 03, the passage "*empty of people running*" is awkward, it would be better to write "*no one is running in the flower field*").

5.2. Comparing Models

We trained NLI classifiers using the train and validation sets of InferBR, SickBR, and ASSIN2. The resulting classification models were used to predict instances in three settings: (i) *intra-dataset* – the predictions are made on the test set of the original dataset; (ii) *inter-dataset* – the predictions are made on the test set of a different dataset; and (iii) *inter***-dataset* – the predictions are made on a different dataset, considering the full set of instances. The goal of the *inter* scenarios is to test the generalization power of the models trained on each

01) **P: O gato está miando no megafone** *The cat is meowing into the megaphone* **Source of error**: premise generation process B using sentence (O) from SICK-BR: **O**: O papagaio está falando no microfone *The parrot is speaking into the microphone*
02) **P: Um homem e uma mulher estão um de frente para o outro, com pássaros circulando as ramificações que emergem de suas cabeças.** *A man and a woman face each other, with birds circling the branches that emerge from their heads.* **Source of error**: Not an real image **O:** Casal, homem e mulher, um de frente ao outro. De suas cabeças saem ramificações com pássaros em volta. *Couple, man and woman, facing each other. Branches come out from their heads with birds around them.*
03) **H: O campo florido está vazio de pessoas correndo.** *The flower field is empty of people running.* **Source of error**: The entailment hypothesis is not well-written based on a well-written premise. **O:** Não há ninguém correndo livremente em um campo florido. *There's no one running free in a field of flowers*

Table 7: Examples of instances with confusing text and the investigated reason. In italics, the sentences are translated to English.

dataset. To do that, the models are presented with sentences that were not previously seen and that may come from a different domain. With that in mind, we did not run the model trained on SICK-BR to predict ASSIN2 labels, and vice-versa, because the overlap of the two datasets is very high, as can be seen in Table 4.

For each dataset, we finetuned the Portuguese BERT model BERTimbau (Souza et al., 2020) for eight epochs using early stopping criteria (if validation loss stops decreasing after three steps – each step was configured to be half an epoch). We kept the same hyperparameters in all models, namely learning rate of 3e-05, dropout of 0.1, and used the AdamW (Loshchilov and Hutter, 2017) optimizer.

Table 8 presents the average classification results in terms of Accuracy and macro-F1 averaged across ten runs with different seeds. Each trained model was tested on the three scenarios. The intra-dataset performance (lines 1 and 2), as expected, always outperforms the inter scenarios. Comparing lines 3 and 4, we notice that the model trained on InferBR generalized better to SICK-BR than the reverse (line 4 has higher accuracy and macro-F1 than line 3). The same pattern repeats in lines 5 and 6. Analyzing the averaged F1-score for each class, we see that models have difficulties recognizing contradictions in the inter-dataset settings (*i.e.,* the scores in column C-F1 are always lower in lines 3 to 6 compared to lines 1 and 2).

	Train	Test	Acc	F1ma	C-F1	E-F1	N-F1
1	SICK-BR	SICK-BR	.85	.85	.86	.81	.87
2	InferBR	InferBR	.90	.90	.89	.91	.90
3	SICK-BR	InferBR	.60	.58	.46	.64	.64
4	InferBR	SICK-BR	.64	.63	.54	.71	.65
5	SICK-BR	InferBR*	.59	.57	.44	.64	.64
6	InferBR	SICK-BR*	.65	.64	.56	.70	.65

Table 8: Classification results considering three classes: (C)ontradiction, (E)ntailment, and (N)eutral. The symbol * means prediction over the entire dataset.

We also ran experiments with two classes (entailment and non-entailment). We transformed the contradiction and neutral instances from SICK-BR and InferBR to non-entailment. The results are in Table 9. Again, InferBR achieved superior accuracy and macro-F1 in all inter-dataset scenarios (lines 4 to 11).

	Train	Test	Acc	F1ma	NE-F1	E-F1
1	ASSIN	ASSIN	.87	.87	.86	.88
2	SICK-BR	SICK-BR	.88	.86	.91	.80
3	InferBR	InferBR	.93	.92	.95	.90
4	ASSIN	InferBR	.76	.72	.83	.62
5	InferBR	ASSIN	.82	.82	.82	.82
6	ASSIN	InferBR*	.77	.73	.83	.62
7	InferBR	ASSIN*	.80	.80	.79	.80
8	SICK-BR	InferBR	.78	.72	.85	.60
9	InferBR	SICK-BR	.79	.76	.84	.68
10	SICK-BR	InferBR*	.78	.72	.85	.58
11	InferBR	SICK-BR*	.79	.76	.84	.67

Table 9: Average classification results from ten models for each dataset with two classes: entailment (E) and Not-Entailment (NE). The symbol * means prediction over the entire dataset.

6. Conclusion

This work produced an NLI dataset for Portuguese using a semiautomatic process to generate premises and an automatic process to generate hypotheses. The processes we used here can be easily adapted to other languages and tasks.

We were the first to use a challenging dataset in Portuguese with image captions, PraCegoVer, with single detailed descriptions per image. It has some advertising images, sometimes unreal, that can produce confusing premises. It was also necessary to minimize the occurrence of named entities and modify 15% of the premises with specific characteristics. We also innovated in the way we generate the premises using only 15% of an existing NLI dataset and producing very different sentences within a similar context.

We used language models for all the automatic transformations: summarization, named entity removal, and text generation. For the generative tasks, we used the paid API of OpenAI using the GPT-4 chat endpoint. From the total amount spent, 85% was to generate the hypotheses and 15% for the premises.

The manual validation concluded that labels were correctly assigned in almost all cases, and only 2.6% of the instances had issues with quality. Models trained on InferBR were better at recognizing entailment in the other Portuguese datasets than the other way around. This may indicate a better generalization power.

In future work, we plan to apply generative models that are open source and compare results. We may expand our dataset using other text-generation strategies and augmentation procedures. We may go more profoundly into the surface properties of the entailment relations of the dataset, comparing them to other datasets from other languages, for example. We also will invest in producing resources for other NLP tasks, adapting the process described here to meet the task objectives.

7. Acknowledgements

This work has been partially funded by CENPES Petrobras, CNPq-Brazil, and Capes Finance Code 001.

8. Ethical Considerations and Limitations

Our work has limitations, which we aim to address in our future work. The number of annotators we used was small but specialized. However, we recognize that once we expand our dataset, more reviewers will be needed in the validation. If the size of the dataset becomes too large, sampling techniques will be applied, similar to what was done in (Bowman et al., 2015b). The reviewers who volunteered to participate in this project could annotate labels at a reasonable pace and always had an open channel to send questions or concerns.

For premise generation according to Process A, we limited our work to an image-captioning dataset. Nevertheless, we believe the process can be easily adapted to generate premises from existing unlabeled corpora, adding more diversity in terms of genre to be more representative of the country's culture.

The ethical considerations regarding this paper are related to using LLM for text generation, especially GPT-4, which does not release information on its training data. We plan to evaluate and adapt our processes to use open-source LLM, which can handle Portuguese. And depending on the quality we get, we also plan to adapt LLMs to enable them to follow instructions and deal better with Portuguese.

All data we used as a source for new premises or train NLI models are public, and no private data is involved. All automatically generated sentences were manually checked, and although nothing caught the attention of the reviewers, they may still contain societal biases, for example, a prevalence of instances with specific race, ethnicity, gender, age, religion, abilities, socioeconomic profile, etc. This may be part of future work, not depending exclusively on reviewers' judgment but also using existing models and tools that can help identify and mitigate those kinds of biases.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sushma Anand Akoju, Robert Vacareanu, Eduardo Blanco, Haris Riaz, and Mihai Surdeanu. 2023. Synthetic dataset for evaluating complex compositional knowledge for natural language inference. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 157–168, Toronto, Canada. Association for Computational Linguistics.

Markus Bayer, Tobias Frey, and Christian Reuter. 2023. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *Computers & Security*, 134.

Luciana Bencke and Viviane Pereira Moreira. 2023. Data augmentation strategies to improve text classification: a use case in smart cities. *Language Resources and Evaluation*, pages 1–36.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Felipe de Souza Salvatore, Marcelo Finger, Roberto Hirata, and Alexandre G Patriota. 2023. A resampling-based method to evaluate nli models. *Natural Language Engineering*, pages 1–28.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2022. # pracegover: A large dataset for image captioning in portuguese. *Data*, 7(2):13.

Daniel Jurafsky and James H Martin. 2023. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Jerrold J. Katz. 1972. *Semantic Theory*. Harper & Row, New York,.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics.

Louise McNally. 2021. Existential.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 406–412. Springer.

Livy Real, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruna Thalenberg, Bruno Guide, Cindy Silva, Guilherme de Oliveira Lima, Igor CS Câmara, Miloš Stanojević, et al. 2018. SICK-BR: a portuguese corpus for inference. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 303–312. Springer.

Mobashir Sadat and Cornelia Caragea. 2022. Learning to infer from unlabeled data: A semi-supervised learning approach for robust natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Johan Van Benthem. 2008. A brief history of natural logic.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014a. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014b. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Zhaorui Zhu, Hongyi Yu, Caiyao Shen, Jianping Du, Zhixiang Shen, and Zhenyu Wang. 2023. Causal language model aided sequential decoding with natural redundancy. *IEEE Transactions on Communications*.