Negation-Aware Data Augmentation for Portuguese Natural Language Inference

Maria Cecília M. Corrêa[1], Felipe S. F. Paula[1], Matheus Westhelle[1], Viviane P. Moreira[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)

{mcmcorrea,fsfpaula,mwesthelle,viviane}@inf.ufrgs.br,

Abstract

Negation plays a fundamental role in human communication and logical reasoning, yet it remains underrepresented in natural language inference (NLI) datasets. This work investigates the impact of targeted data augmentation using negation cues on the main NLI datasets for Portuguese (INFERBR, ASSIN, and ASSIN 2). By synthetically generating new instances with negated hypotheses, we create more diverse training and test sets. A BERT-based model was fine-tuned and tested on the combined datasets and augmented data. The results show that the model was heavily influenced by the bias in the use of negation, and increased data diversity improves the model's handling of negation.

1 Introduction

Negations are a key linguistic phenomenon that allows speakers to describe properties that people or things do not hold or events that do not happen (Jiménez-Zafra et al., 2020). However, despite its ubiquity, negation remains underrepresented in the training data available for various language tasks. Negation has been shown to be challenging even for humans to correctly interpret objectively due to the diversity of forms across fields (Truong et al., 2022) and in combination with other linguistic phenomena. For instance, complex expressions such as gradable adjectives (*e.g.,* not uncommon does not imply common) and downward entailment ("*A man owns a dog*" entails "*A man owns an animal*" but "*A man does not own a dog*" does not entail "*A man does not own an animal*") can be harder to detect and reason.

Natural language inference (NLI) is the task of determining the logical relationship between a premise and a hypothesis. In this context, the ability to correctly interpret negated statements is not only desirable but necessary. Previous studies (Hossain et al., 2020) have shown that even large pre-trained language models often misclassify negated inputs, especially when such patterns are underrepresented in the training data. However, the vast majority of research and resources for this task are concentrated in English. Since linguistic phenomena like negation are language-specific, findings from English cannot be directly generalized, highlighting the critical need for dedicated benchmarks and studies in other languages, such as Portuguese.

Constructing new NLI datasets to capture specific linguistic properties is often a complex task (Sadat and Caragea, 2022). Consequently, research that reduces dependence on costly manual annotation for training NLI models has become essential. Data augmentation is a decisive technique for developing robust benchmarks, as it addresses the problem of data scarcity for specific linguistic phenomena. For negation, this means strategically generating new, varied synthetic instances. By enriching the training distribution with such curated examples, models can learn to distinguish meaningful logical relationships from superficial cues, thereby improving their capacity to represent and reason with negation.

This work addresses the impact of targeted data augmentation using negation cues in the context of three large-scale NLI datasets for Portuguese, namely INFERBR (Bencke et al., 2024), ASSIN (Fonseca et al., 2016), and ASSIN 2 (Real et al., 2020). We generated synthetic hypothesis statements through controlled negation, producing additional training and test instances with diverse semantic relationships. The augmentation process was guided by large language models (LLMs), which helped construct meaningful negated variants that preserve fluency and contextual relevance. The result is a more robust dataset that challenges the ability of the models to distinguish fine-grained logical relations.

To evaluate the effects of the augmentation,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 141–150

we fine-tuned the Portuguese BERT-based model BERTimbau (Souza et al., 2020) on the original instances of the combined datasets and on our new negated augmented ones [1], comparing the performance in the face of different types of negation. Our experiments revealed that model predictions are significantly affected by the presence of negation, particularly syntactic ones. The model trained on the augmented dataset was able to maintain the performance achieved by the one trained only on original data while also improving its reasoning of *entailment* instances in the presence of negation cues. These findings emphasize the importance of incorporating negation diversity into NLI benchmarks and expanding it to languages other than English, as negation is a language-dependent phenomenon.

The contributions of this paper are (i) a study of the impact of negation in Portuguese NLI, (ii) a manually curated NLI dataset created by augmenting existing resources to produce a more balanced representation of negation, (iii) an NLI model capable of making more accurate inferences in the presence of negation. By demonstrating both the benefits and limits of negation-based augmentation, this study contributes insights into dataset construction and model evaluation strategies for semantic reasoning tasks in Portuguese.

2 Background and Related Work

Natural Language Inference (NLI), also known as recognizing textual entailment (RTE), consists of inferring from a premise whether a given hypothesis is considered to be true. It can be presented as a binary classification problem, considering the labels *entailment* and *not entailment*, but many NLI datasets, such as Stanford Natural Language Inference (Bowman et al., 2015), are labeled with three classes (*entailment*, *contradiction*, and *neutral*). Taking the premise *All fifteen guests were in the kitchen* as an example, it entails the hypothesis *There were no guests in the bedroom*, contradicts *The house was empty*, and is neutral towards *It was a sunny day*.

Many downstream NLP tasks benefit from NLI mechanisms, such as information retrieval (Dai et al., 2024), question answering (Yao and Barbosa, 2024), and text summarization (Maynez et al., 2020; Sridhar and Visser, 2023). Widely adopted benchmarks include the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018), which offer large-scale entailment, contradiction, and neutral labeled pairs in English. For Portuguese, the ASSIN datasets (Fonseca et al., 2016; Real et al., 2020) provide natural inference pairs annotated for both semantic similarity and entailment, while INFERBR (Bencke et al., 2024) introduces a large-scale machine-generated with verification NLI resource.

Data Augmentation is a technique used to increase the diversity and volume of training data without the need for additional manual annotation (Feng et al., 2021; Chen et al., 2023; Bencke and Moreira, 2024). Augmentation helps mitigate overfitting, improves generalization to unseen data, and is particularly valuable in low-resource scenarios or when datasets are imbalanced across classes. The emergence of large language models (LLMs) such as GPT-4 (OpenAI et al., 2024) has significantly advanced the capabilities of data augmentation. GPT-4 can generate contextually appropriate and grammatically correct variations of existing sentences, introducing specific linguistic features with high fluency. Through the OpenAI API, users can programmatically submit prompts and retrieve model-generated outputs at scale.

Negation Representation. Despite its pervasiveness, negation remains a challenging and underrepresented aspect in many NLP benchmarks. A survey of corpora annotated with negation highlights the diversity of negation types and the challenges in consistently capturing them across languages and tasks (Jiménez-Zafra et al., 2020), which directly impact the performance of language understanding systems, particularly in tasks that require nuanced reasoning. Analyses of NLI datasets reveal that negation is not only underrepresented, but also often not sufficiently varied, leading to poor generalization by models when faced with negated inputs (Hossain et al., 2020). This shortcoming is further demonstrated by evaluations showing that LLMs often misinterpret or ignore negation in benchmark tests (Truong et al., 2023). The problem is not limited to text-based models: in the multi-modal domain, vision-language systems also benefit from explicit negation learning. Recent approaches use LLMs to generate synthetic,

[1] All augmented instances, along with a compilation of found negation cues, dataset distributions, and other resources, can be found in the project repository https://github.com/cixcore/neg-nli

negation-rich image-text pairs (Singh et al., 2024; Alhamoud et al., 2025). These instances highlight common failures in handling negated content, and when incorporated into training, significantly enhance model sensitivity to negation and robustness in reasoning tasks.

NLI for Portuguese. Despite its key role in natural language understanding, most NLI benchmarks are centered on the English language. For Portuguese, the ASSIN (Fonseca et al., 2016) dataset was the first large-scale NLI resource, providing sentence pairs extracted from news articles written in European and Brazilian Portuguese and annotated both with similarity scores and with entailment, paraphrase, or neutral labels. Expanding on the task, ASSIN 2 (Real et al., 2020) introduced a benchmark built upon SICK-BR (Real et al., 2018). While SICK-BR is a translated adaptation of the English SICK corpus (Marelli et al., 2014), ASSIN 2 extended this foundation by focusing on data balancing to address potential class distribution biases and create a more reliable evaluation resource for Portuguese NLI. Lastly, addressing the problems of syntactic variation and inference complexity of the previous contributions, the INFERBR dataset (Bencke et al., 2024) provided a large-scale, manually curated benchmark specifically for Brazilian Portuguese. To construct the dataset, the authors used LLMs to generate diverse and semantically rich hypotheses, which were then curated to ensure linguistic naturalness, unambiguity, and label reliability. An inter-dataset evaluation showed that models trained on ASSIN or SICK-BR performed significantly worse when evaluated on INFERBR, indicating poor generalization. This finding highlights the unique challenges posed by INFERBR and reinforces the need for more comprehensive and varied NLI datasets in Portuguese to support the development of robust inference models.

NLI datasets that are currently available exhibit significant bias due to insufficient negated data and unbalanced representation. In this work, we aim to assess model sensitivity to negation cues in Portuguese and introduce a data augmentation framework that explicitly generates negated examples on underrepresented scenarios. Our approach provides a clear path to more robust and generalizable Portuguese language models, mitigating negation-related bias.

3 Negation representation in Portuguese

In this section, we cover the datasets used and their normalization; our approach to negation cue identification through manual curation and semi-automated analysis; and negation distribution results showing hypothesis negations strongly used to represent "not entailment" instances.

3.1 Datasets

The Portuguese NLI resources picked for negation analysis were the ASSIN (Fonseca et al., 2016), ASSIN 2 (Real et al., 2020) (which contains SICK-BR), and INFERBR (Bencke et al., 2024) datasets. Both ASSIN and ASSIN 2 datasets do not distinguish between neutral and contradiction relationships, so INFERBR was normalized by mapping neutral and contradiction labels as *not entailment*. ASSIN also has a paraphrase label that was mapped as *entailment*. This process resulted in a combined binary dataset unbalanced towards *not entailment* relationships, as shown in Table 1. We will refer to the original combined training and test sets as $train$ and $test$ respectively, and the subset of instances from $test$ containing any type of negation on the hypothesis as $testNeg$.

Dataset	Not entailment	Entailment	Sum
ASSIN	7316 (73.1%)	2684 (26.8%)	10000 (33.0%)
ASSIN 2	4724 (50.0%)	4724 (50.0%)	9448 (31.2%)
INFERBR	7202 (66.6%)	3601 (33.3%)	10803 (35.7%)
Total	19242 (63.6%)	11009 (36.3%)	30251

Table 1: Distribution of instances per label after combining all datasets

3.2 Negation Cues

To establish the foundation of our negation analysis, we compiled a non-exhaustive list of common Portuguese negation markers. Previous work on negation annotation highlighted three types of negation cues (Jiménez-Zafra et al., 2020): **syntactic**, which involve independent grammatical markers like *não* (*no/not*) or *nunca* (*never*); **lexical**, where the negation is embedded in the meaning of the expression itself, such as *negar* (*to deny*), *evitar* (*to avoid*) or *privar de* (*deprive of*); and **morphological**, where negation is expressed through a morpheme, notably prefixes in Portuguese, *e.g., in-* in *incoerente* (*incoherent*), *des-* in *desfazer* (*undo*).

The candidates for syntactic and lexical cues were manually proposed and curated, removing expressions that caused disagreement between the

annotators. Morphological expressions followed a semi-automated process based on Portuguese negation prefixes (*e.g., a-, des-, dis-, i-*). For each prefix, we identified words that, if removed, would result in another valid word. However, there are cases where that is undesirable, for instance, in the word "*incenso*" ("*incense*"), where removing the "*in-*" prefix would result in "*censo*" ("*census*"), a different lexeme altogether; to avoid this, we filtered words using an LLM, keeping only cases where adding a prefix genuinely results in a morphologically derived word. All cues were then lemmatized, except for lexical negation expressions that were considered as-is. The result was a list with 2059 categorized negation cues. A sample is presented in Table 2, and the complete set of instances is available in our repository.

Token	Type	Lemma
exceto (except)	Syntactic	exceto
salvo (saved)	Syntactic	salvo
não (no)	Syntactic	não
nenhum (none)	Syntactic	nenhum
ninguém (nobody)	Syntactic	ninguém
nunca (never)	Syntactic	nunca
jamais (never)	Syntactic	jamais
nada (nothing)	Syntactic	nada
sem (without)	Syntactic	sem
sequer (not even)	Syntactic	sequer
nem (nor)	Syntactic	nem
tampouco (neither)	Syntactic	tampouco
falta de (lack of)	Lexical	EXPRESSION
de modo algum (in no way)	Lexical	EXPRESSION
sem chance (no chance)	Lexical	EXPRESSION
de maneira nenhuma (no way)	Lexical	EXPRESSION
de forma alguma (in no way)	Lexical	EXPRESSION
privado de (deprived of)	Lexical	EXPRESSION
evitar (avoid)	Lexical	evitar
ausência (absence)	Lexical	ausência
ausente (absent)	Lexical	ausente
cessar (cease)	Lexical	cessar
negar (to deny)	Lexical	negar
recusar (refuse)	Lexical	recusar
rejeitar (reject)	Lexical	rejeitar
opor (oppose)	Lexical	opor
resistir (resist)	Lexical	resistir
parar (to stop)	Lexical	parar
excluir (delete)	Lexical	excluir
proibir (prohibit)	Lexical	proibir
desinteressado (disinterested)	Morphological	desinteressar
desinteressante (uninteresting)	Morphological	desinteressante
desinteressar (disinterest)	Morphological	desinteressar

Table 2: Sample of negation cues categorized by type.

3.3 Negation Distribution

To understand how negation is represented in all three datasets, all instances were lemmatized and searched for the lemmas extracted as described in Section 3.2, with lexical expressions searched separately without lemmatization. Each dataset individually contains roughly 20% of instances with negation in either the premise or hypothesis.

A significant bias is evident: when negation appears in a hypothesis, it strongly predicts a contradiction or non-entailment label. For most cues, this association exceeds 76.37%, the lowest rate, found in the ASSIN training set and higher than the dataset's inherent class imbalance. This reflects a known phenomenon in NLI datasets, where negation is frequently introduced to obtain contradictions (Gururangan et al., 2018). After combining all datasets, 84.37% of train and 81.76% of test instances with some form of negation in the hypothesis are classified as not entailment, and the most present type is syntactic negations, as shown in Tables 3 and 4.

	train	*test*
% Instances with Negation	20.80	20.86
% Premises with Negation		
Total	10.30	11.49
Entailment	23.61	22.91
Not entailment	**76.39**	**77.09**
% Hypotheses with Negation		
Total	12.90	12.49
Entailment	15.63	18.24
Not entailment	**84.37**	**81.76**

Table 3: Distribution of negation cues in *train* and *test* splits of all datasets combined

Type	Count	Entailment	Not entailment
Lexical	142	21.8%	78.17%
Morphological	354	27.6%	72.32%
Syntactic	**4061**	**12.6%**	**87.3%**
Total	4557	643 (14.1%)	3914 (85.9%)

Table 4: Negation types distribution in the combined dataset

4 Generating Negated Hypotheses

Figure 1 outlines the augmentation process, which is described as follows.

1. **Filter instances without negation cues.** The collection of negation cues mentioned in Section 3.2 was used to filter instances that had no negation present in their hypotheses. We specifically picked hypotheses on the basis that the nature of negations is to represent the absence of concepts, and a negated premise

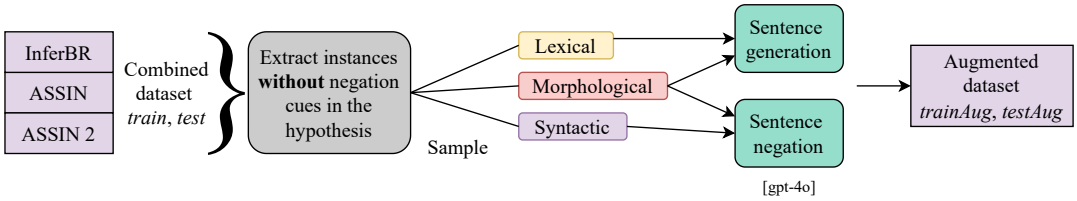


Figure 1: Augmentation pipeline

had a tendency to generate less intuitive relationships, while a negated hypothesis would have an easier context to reason. We also selected only instances labeled *not entailment*, taking into consideration that the label was overrepresented and our goal was to increase the presence of *entailment* (which is presumably the negation of a *not entailment*).

2. **Sentence negation and automatic correction.** We used GPT-4o to propose instances negating the hypotheses. The original training instances generated augmented training instances, and the test instances were likewise augmented, to avoid train-test contamination. Each negation type had a different generation approach:

    - *Lexical* negations were obtained straight from the candidates list, where each expression was used to generate 20 premise-hypothesis pairs from scratch varying nominal and verbal inflection. An example result of this approach would be: **premise:** "*O professor deixou claro que colar na prova resulta em expulsão*" ("*The teacher made it clear that cheating on the exam results in expulsion*"); **hypothesis:** *Colar é de modo algum tolerado* ("*Cheating in the exam is by no means tolerated*").
    - *Morphological* negations used a hybrid approach, where most instances were generated from scratch, similarly to the lexical procedure, and the rest used a sample from each dataset that contained any of the morphological candidates without a prefix. The sub-word was then negated by adding a negation prefix; however, this did not produce high-quality results, as the negation of the word did not consistently match its context. For instance, the morphological negation of the sentence "*Uma menina está usando uma faca para cortar o bolo*" (*A girl is using a knife to cut the cake*) would be "*Uma menina está desusando uma faca para cortar o bolo*", (*A girl is unusing a knife to cut the cake*), which does not make sense. The instances that produced sound sentences were combined with the ones that followed the lexical procedure.
    - *Syntactic* negations were generated using a three-step procedure. First, a syntactic cue was applied to 300 original instances sampled from each dataset split, totaling 1800 instances (900 for training and 900 for testing). Then, the instances were revised to ensure they adhered to the *entailment* relationship (*i.e.,* replacing a word with its antonym, changing an object for one not mentioned in the premise, adding existence quantifiers, *etc.*). Finally, we executed an automatic correction step to improve instances with incorrect labels or confusing sentences.

3. **Manual verification.** The quality and correctness of the augmented instances underwent human validation. Instances with incorrect labels, confusing sentences, or without an effectively introduced negation were discarded. We also discarded instances in which the hypothesis was an exact copy of the premise or leaked the reasoning of the LLM (*e.g.,* "*The ball was blue*" negation turning up as "*It can not be inferred from the premise that the ball was blue*").

By the end of the process, we were left with 794 synthetic test instances and 795 training instances. The synthetic instances are concatenated to the original sets $train$ and $test$, producing the new augmented datasets $trainAug$ and $testAug$. Table 5 presents the total of each augmented type.

The original distribution of negation cues is concentrated on syntactic negations, so the augmentation process focused on that type.

The manual revision process excluded over a third of each syntactic set. The *not entailment* relations were frequently too neutral, holding information that was hard for the LLM to manipulate into a negated entailment, like a hypothesis with almost no relation to the premise or with information that was not mentioned, and even different narrative forms between premise and hypothesis.

Type	Train	Test
Lexical	180	180
Morphological	56	31
Syntactic	559	584
Total	795	794

Table 5: Augmented instances by negation type for combined train and test sets

5 Experimental Evaluation

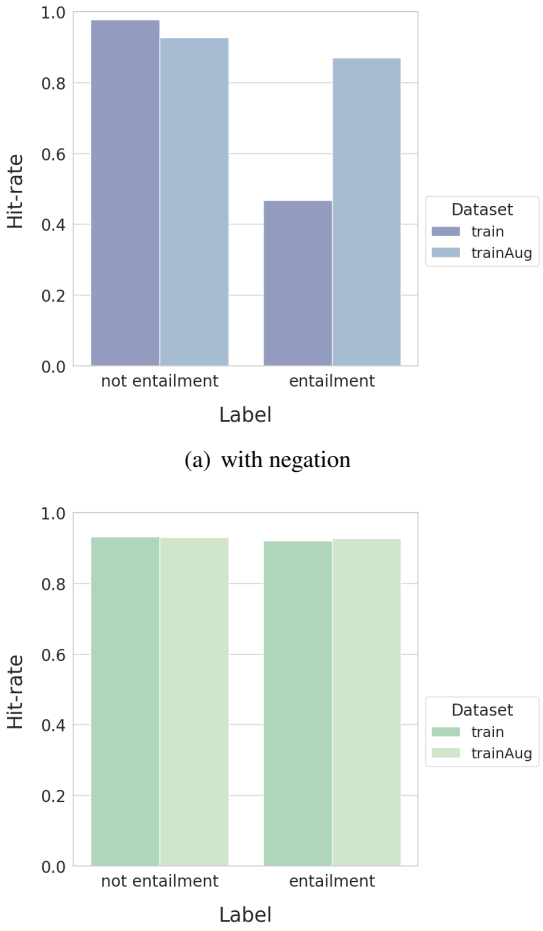
This section evaluates whether targeted data augmentation for unbalanced negation representation can mitigate negation-related bias in Portuguese NLI datasets. Using a fine-tuned BERT-based model on original ($train$) and augmented ($trainAug$) training sets, we measure robustness across test sets with varying negation distributions ($test$, $testAug$, $testNeg$).

We reproduced the evaluation performed on INFERBR by fine-tuning the BERTimbau (Souza et al., 2020) model on our two training sets, $train$ and $trainAug$. The same hyperparameters were used for both training sets, namely eight epochs using early stopping criteria, learning rate of 3e-05, a dropout of 0.1, and the AdamW optimizer (Loshchilov and Hutter, 2019).

Dataset	$test$ F_1	$testAug$ F_1
$train$	0.93	0.88
$trainAug$	0.93	0.92

Table 6: Weighted F1-score by training and test sets

While all models achieved accuracy above 90% on dataset $test$, the model trained only on the original set $train$ suffered a significant drop in performance when tested on dataset $testAug$, despite the limited number of new instances. The data bias toward over-representing negation in *not entailment* led to its adoption of this label as a shortcut in the presence of negation cues, regardless of the true meaning of the sentence. This yielded a very low score when *entailment* was the target label as presented in Figure 2.



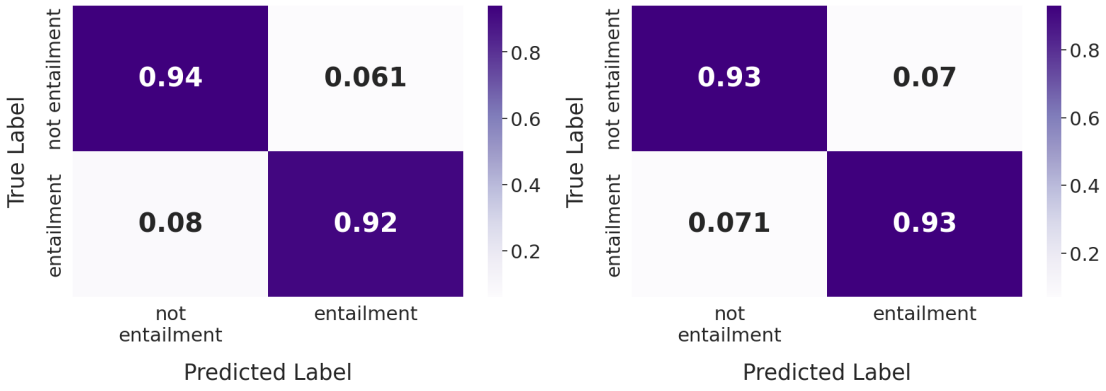
(a) with negation

(b) without negation

Figure 2: Hit-rate by presence of negation on $train$ predicting $testAug$

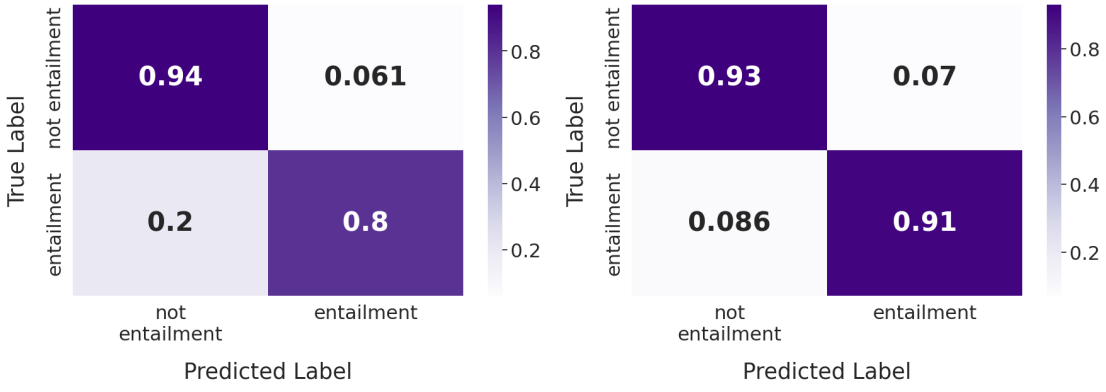
The model trained on $trainAug$ was able to almost maintain the performance achieved on $test$ for $testAug$, showing that the diversity introduced by the synthetic data improved the comprehension of negation, despite only representing 4% of the total training set (Table 6). Detailed confusion matrices are shown in Figures 3 and 4. When evaluating the models on the $testNeg$ subset, the performance achieved by $train$ was superior overall. However, a considerably higher recall was demonstrated by $testAug$, as shown in Figure 5. It highlights how $testAug$ was able to generalize negated instances and not predict *not entailment* as a shortcut, despite suffering penalties from dataset unbalancing.

Syntactic negations were the most affected,



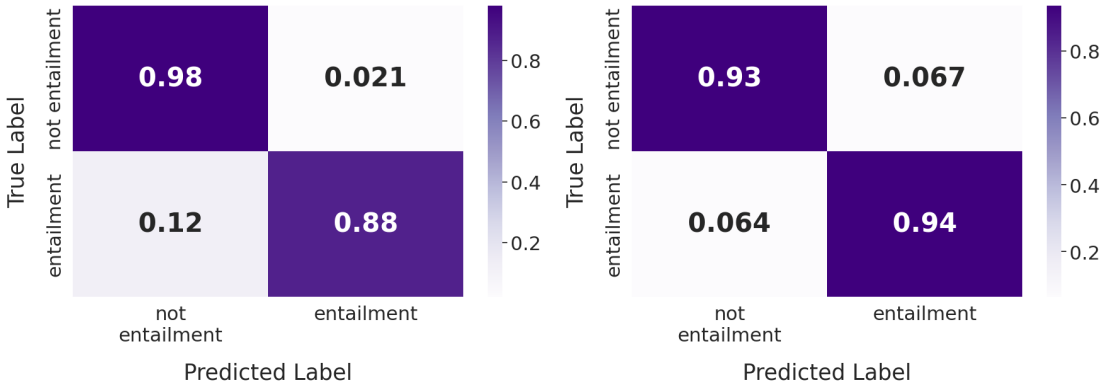
(a) $test$ (b) $testAug$

Figure 3: Confusion Matrices training only with $train$ and predicting $test$ and $testAug$ sets



(a) $test$ (b) $testAug$

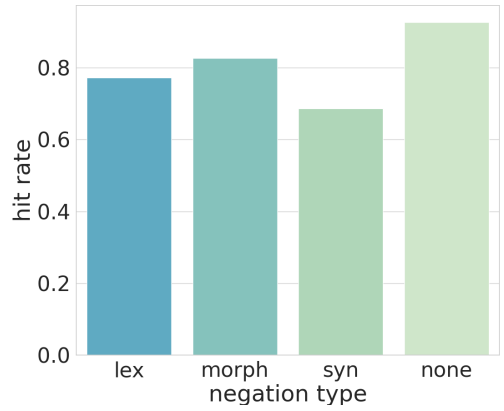
Figure 4: Confusion Matrices training with $trainAug$ and predicting $test$ and $testAug$ sets



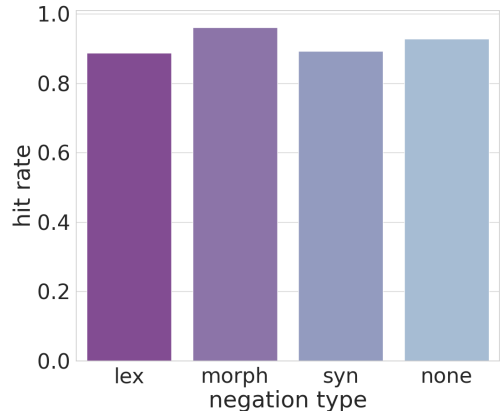
(a) $train$ (b) $trainAug$

Figure 5: Confusion Matrices training with $train$ and $trainAug$ to predict $testNeg$

which was somewhat expected considering they had the greatest unbalancing in distribution, and also had the most instances synthetically generated. For instance, after being exposed to the augmented data, the model was able to predict correctly that the premise "*Uma mulher está pondo de lado um limão*" ("A woman is putting a lime aside") entails the hypothesis "*Uma mulher não está espremendo um limão*" ("A woman is not squeezing a lime"). However, Figure 6 also shows a notable improvement for morphological negations, with $trainAug$ being able to predict all the new test instances cor-



(a) $train$



(b) $trainAug$

Figure 6: Hit-rate by negation type on the $testAug$ set

rectly for that category.

From a total of 89 augmented test instances that both $train$ and $trainAug$ trained models predicted wrong, 53 (59%) were generated from ASSIN. This could be because instances from ASSIN tended to have longer sentences with more complex contexts, for example, the pair "*O Arsenal é uma das duas equipas que conseguiu derrotar o Leicester está temporada na Premier League (2-5).*" [*sic*] ("Arsenal is one of only two teams that managed to beat Leicester this season in the Premier League (2-5).") entailing "*Com este resultado, o Leicester não foi capaz de derrotar o Arsenal.*" ("With this result, Leicester were unable to defeat Arsenal.").

6 Conclusion

This work investigated the role of negation in natural language inference for Portuguese by applying targeted data augmentation to the ASSIN, ASSIN 2, and INFERBR datasets. We used a large language model to synthetically generate negated instances for three negation types classified as the under-represented *entailment* class. Our experiments demonstrated that models trained only on the combined original dataset exhibited significant performance drops when evaluated on test sets enriched with negation. This suggests a strong reliance on shallow lexical cues, particularly associating negation with *not entailment*, regardless of semantic context. Introducing synthetic negated examples enhanced the models' ability to generalize, resulting in more accurate predictions in the presence of negation, even with limited augmented data available.

Future work will explore a more robust augmentation process for generating instances with different negation types, rather than focusing solely on syntactic negations. Most Portuguese NLI resources provide only binary (*entailment* / *not entailment*) instances. Among the datasets used, INFERBR was the only one that distinguished between neutral and contradiction relations. This distinction was collapsed during preprocessing to ensure a uniform data augmentation and evaluation methodology could be applied across all datasets. A clear improvement would be to tailor the augmentation methods to preserve and leverage this ternary classification, preventing the loss of nuanced information. Another direction involves probing negation importance, *i.e.,* whether negation meaningfully changes the value of a statement. There are also other challenges associated with negation and NLI that require investigation, such as quantification (*e.g.,* "not all," "almost no"), hedging ("unlikely," "doubtful"), lexical relations (morphological negation, antonymy), and downward-entailment.

7 Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and CNPq.

References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. *Preprint*, arXiv:2501.09425.

Luciana Bencke and Viviane Pereira Moreira. 2024. Data augmentation strategies to improve text clas-

sification: a use case in smart cities. *Language Resources and Evaluation*, 58(2):659–694.

Luciana Bencke, Francielle Vasconcellos Pereira, Moniele Kunrath Santos, and Viviane Moreira. 2024. InferBR: A natural language inference dataset in Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9050–9060, Torino, Italy. ELRA and ICCL.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Lu Dai, Hao Liu, and Hui Xiong. 2024. Improve dense passage retrieval with entailment tuning. *Preprint*, arXiv:2410.15801.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *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)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, 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 C. S. Câmara, Miloš Stanojević, Rodrigo Souza, and Valeria de Paiva. 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*, page 303–312, Berlin, Heidelberg. Springer-Verlag.

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*, pages 4763–4776, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn "no" to say "yes" better: Improving vision-language models via negations. *Preprint*, arXiv:2403.20312.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Arvind Krishna Sridhar and Erik Visser. 2023. Improved beam search for hallucination mitigation in abstractive summarization. *Preprint*, arXiv:2212.02712.

Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. Improving negation detection with negation-focused pre-training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4188–4193, Seattle, United States. Association for Computational Linguistics.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *Preprint*, arXiv:2306.08189.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Peiran Yao and Denilson Barbosa. 2024. Accurate and nuanced open-qa evaluation through textual entailment. *Preprint*, arXiv:2405.16702.