How aspects of similar datasets can impact distributional models*

Isabella Maria Alonso Gomes[1], **Norton Trevisan Roman**[1]

[1]School of Arts, Sciences and Humanities – University of São Paulo (USP)

{isabellaagomes,norton}@usp.br

Abstract. *Distributional models have become popular due to the abstractions that allowed their immediate use, with good results and little implementation effort when compared to precursor models. Given their presumed high level of generalization it would be expected that good and similar results would be found in data sets sharing the same nature and purpose. However, this is not always the case. In this work, we present the results of the application of BERTimbau in two related data sets, built for the task of Semantic Similarity identification, with the goal of detecting redundancy in text. Results showed that there are considerable differences in accuracy between the data sets. We explore aspects of the data sets that could explain why accuracy results are different across them.*

1. Introduction

In recent years, distributional models have become very popular due to the abstractions created that allowed for their immediate use, with good results and relatively little effort by researchers when compared to some of their precursor models. Not surprisingly, it is currently possible to find them applied to problems related to a vast range of tasks, such as question answering (*e.g.* [Yang et al. 2019]), automatic summarization (*e.g.* [Liu 2019]), plagiarism (*e.g.* [Rosu et al. 2021]), among others.

Because they are trained in broad open domain corpora, pre-trained models like BERT (and their derivations) enable implementations in a simplified way, allowing for the configuration of pre-set parameters so as to reflect the specificity of each task [Devlin et al. 2018]. Given their presumed high level of generalization [Hendrycks et al. 2020], it would be expected that good results would be found in different data sets and that fine tuning would suffice to improve application performance without overfitting.

If models like BERT are expected to present good results in data sets from different domains, it would also be expected that in data sets sharing the same nature and serving the same purpose they would not only present good results, but that these results would also be similar. Would it be possible, however, that the application of BERT and its derivatives in similar data sets would present considerably different results? If so, what data set features could justify such outcomes?

To help addressing questions like these, some efforts have already been spent in determining how different models behave in different data distributions . In fact, there

*The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation.

is an indication that a shift in focus from data quantity to data quality could lead to robust models and improve out-of distribution generalization [Swayamdipta et al. 2020]. Besides that, labeling errors have already been found to be responsible for reducing the accuracy of machine learning models, even when these are tested in similar data sets (e.g. [Swayamdipta et al. 2020]).

In this work, we move one step forward towards answering these questions, by presenting the results of the application of BERTimbau [Souza et al. 2020], a variation of BERT for Portuguese, in two related data sets, built for the task of Semantic Similarity identification, with the ultimate goal of detecting redundancy in text. As it will be made clearer in the forthcoming sections, accuracy results were different across data sets, even though they were supposed to be similar, also serving the same overall purpose.

The rest of this article is organised as follows. In Section 2 we present a brief overview of current results on the variability of machine learning methods in different, however related, data sets. Next, in Section 3 we describe the corpora and models used in our experiments, along with the methodology we followed. Results are, in turn, presented and discussed in Section 4. Finally, Section 5 presents our final remarks and directions for future research.

2. Related Work

The ability to generalize is the subject of different researches within the field of NLP and artificial intelligence in general, whose main objective is to ensure the robustness, reliability and security of machine learning models. Problems such as out of distribution detection (OOD), anomaly detection (AD), novelty detection (ND), open set recognition (OSR), and outlier detection (OD) are constantly gaining more notoriety [Yang et al. 2021]. Although distributional models such as BERT achieve good accuracy, there are aspects that make the implementation of such models in similar data sets (e.g. [Hendrycks et al. 2020]) to present relevant differences in accuracy.

In [Hendrycks et al. 2020] the authors try to answer whether the models have a good generalization ability in new distributions. They measure the generalization ability of bag-of-words, ConvNets, LSTMs, and pre-trained transformers' models in out-of-distribution (OOD) examples applied to seven different data sets. They found that pre-trained transformers are more effective at detecting anomalous examples or OOD, while many previous models are often worse than chance. In addition, they analyze which factors affect the robustness of models, finding that models trained on larger bases are not necessarily more robust, with the diversity of data in training being more relevant to ensure robustness.

Results showed that pre-trained Transformers are often more robust, presenting a smaller generalization gap. For the LSTM model, performance dropped by more than 35% whereas for RoBERTa the generalization performance even increased. As an evidence that larger data sets do not always represent better generalization, the use of BERT Large did not reduce the generalization gap. However, having greater data diversity in the pre-trained models was found to improve their generalization, as it was observed a greater robustness in RoBERTa in relation to BERT Large. For the out of distribution detection, pre-trained transformers models also proved to be better detectors. However, the authors suggest that there is still room to improve the out of distribution detection mechanisms

even for this type of model.

In [Tu et al. 2020], the authors proposed to use multi-task learning to improve the generalization of pre-trained models like BERT for inference and paraphrase identification, showing that the generalization improves when the minority class is inflated, suggesting that diversity in the data set may be the explanation behind this result. The models tested were BERT Base, BERT Large, RoBERTa base and RoBERTa Large. The authors observed that in the data set with longer and more complex sentences the models' accuracy reduced. It was also observed that a long fine tuning does not help the model in general, but improves the accuracy in the minority class. In addition, it was observed that the models do not allow extrapolation, that is, removing the minority class did not translate into better accuracy.

The results of [Swayamdipta et al. 2020] indicate that a shift in focus from data quantity to quality can lead to robust models and improve out of distribution generalization. The objective of this research was to map the data set so as to show the presence of ambiguous regions that contribute the most to generalization out of the distribution. In addition, it also maps the most populated regions in the data, which are easy to learn and play an important role in model optimization. Finally, data maps reveal a region with instances that the model found difficult to learn, often corresponding to labeling errors. All experiments were done with ROBERTA Large. After identifying the most populated regions, models were tested exclusively using the examples of each region. The results indicate that training with ambiguous instances promotes generalization, with little or no effect on data distribution. Furthermore, this experiment showed that data sets mostly have easy-to-learn instances, and that hard-to-learn instances were generally related to labeling errors. As it turns out, these results are in line with ours, in that we also found such errors to impact the model's results.

3. Materials and Methods

To explore how aspects of similar data sets can impact the accuracy of the models, we run BERTimbau [Souza et al. 2020] in the ASSIN [Fonseca et al. 2016][1] and ASSIN2 [Real et al. 2020][2] databases, along with a data set we built by grouping together both corpora. The decision to use BERTimbau was due to it is a distributional model tailored to Brazilian Portuguese, and which is based on BERT, a widely used model in many NLP tasks. Both data sets used in the experiments are also in Portuguese.

3.1. Source Corpora

ASSIN was created in 2016, with the objective of providing materials for two tasks: identification of Semantic Similarity and Textual Inference. As such, the corpus comprises pairs of sentences, extracted from news, and written in both European and Brazilian Portuguese. It was compiled through Google News, by selecting similar sentences from different documents, where each document corresponded to the same events. Similarity was calculated through Latent Dirichlet Allocation – LDA. In the sequence, pairs were manually filtered, so as to exclude those that might be considered noisy. News in Brazilian

[1]http://nilc.icmc.usp.br/assin/

[2]https://sites.google.com/view/assin2/

Portuguese were gathered from G1[3], whereas their European Portuguese versions came from Público[4].

Each pair was then manually labeled by four independent annotators regarding both tasks. Six annotation teams took part in the Semantic Similarity task, three from Brazil and three from Portugal, with four teams dealing with Inference. We refer the interested reader to [Fonseca et al. 2016] for details on this procedure. For the semantic similarity score, a continuous scale, ranging from 1 to 5, was used, so as to reflect how similar the content of both sentences in the pair were (Table 1 presents some examples of sentences and their associated scores). Assigned values are:

1. The sentences are completely different. It is possible that they talk about the same fact, but this is not visible by examining them in isolation, without context;
2. The sentences refer to different facts and are not similar to each other, but they are about the same subject (football match, votes, currency variations, accidents, products etc);
3. The sentences have some similarity, and may refer to the same fact;
4. The content of the sentences is very similar, but one (or both) gives some information away which is not present in the other. The difference may be mentioning a different date, place, quantity, or even a different subject or object; and
5. The sentences have pretty much the same meaning, possibly with a slight difference (such as an adjective that doesn't change its interpretation).

Overall, the data set was labeled by 36 people. Of the total number of labeled sentences, 11.3% were discarded because they did not meet the criteria for judgment on textual inference (*i.e.* at least three annotators should reach an agreement). In total, the data set comprises 10,000 pairs of sentences, being half of them written in Brazilian Portuguese and half in European Portuguese. For each annotator, the correlation of their similarity scores with the average scores of the pairs he or she labeled was calculated. The values found show a good agreement between the annotators, with Pearson's correlation = 0.74 for the similarity semantic task.

Finally, the corpus comprises three different data sets (training, validation and test) so as to provide a common ground for comparison across different studies. Figure 1 shows the score distribution in each of ASSIN's data sets. As it turns out, data distribution across sets presents some variation, which may have an impact in the performance of any classificator applied to them. We will come back to this issue later on this section.

ASSIN's second edition, ASSIN2, was also designed for Semantic Similarity and Textual Inference. It was however based on the SICK-BR corpus [Real et al. 2018], a translation and adaptation of the SICK corpus [Marco et al. 2014]. SICK-BR is known for not having complex linguistic phenomena, in which its sentences were generated from simple facts, coming from image captions, and only containing sentences in Brazilian Portuguese. In ASSIN2, all sentence pairs were labeled by at least four native speakers of Brazilian Portuguese with linguistic training, comprising 9,448 sentence pairs. Figure 2 shows the score distribution in each of ASSIN2's data sets. As can be seen in the figure, the differences in data distribution across sets are more pronounced in ASSIN2 when compared to ASSIN.

[3]http://g1.globo.com/

[4]http://publico.pt/

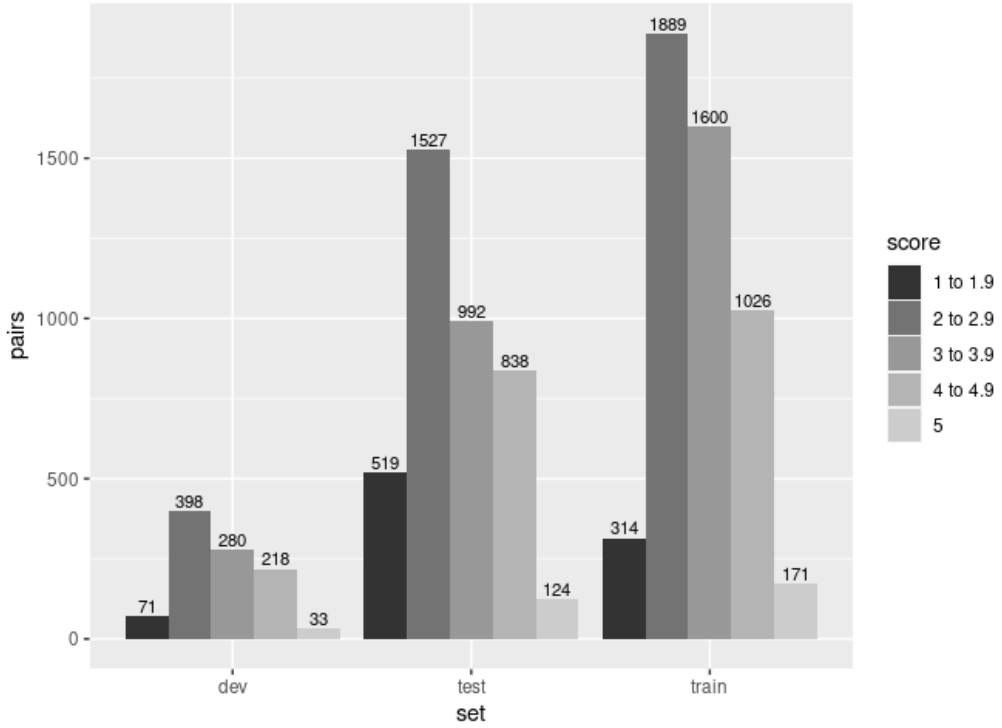

Figure 1. Distribution of scores across data sets in ASSIN.

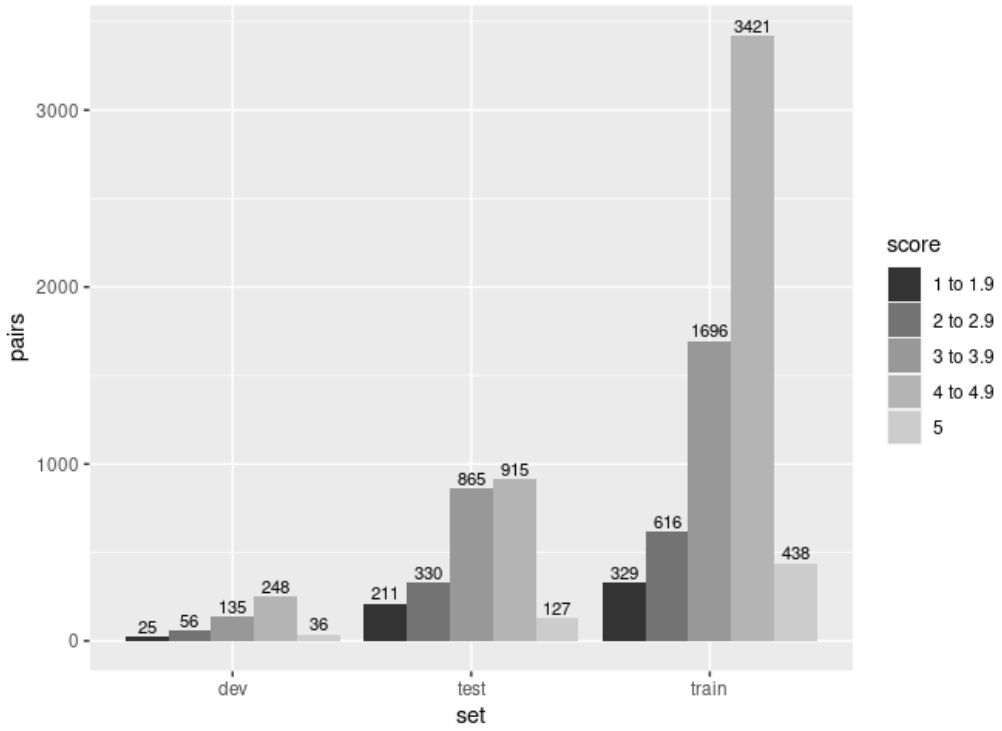

Figure 2. Distribution of scores across data sets in ASSIN2.

Table 1. Sentences and associated similarity scores from ASSIN

Score	Example
1	This is the first time a head of Catholic Church uses the word in public. (*Mas esta é a primeira vez que um chefe da Igreja Católica usa a palavra em público*)
	Yesterday Germany for the first time recognized Armenian genocide. (*A Alemanha reconheceu ontem pela primeira vez o genocídio armênio*)
2	As expected, first half was characterised by balance. (*Como era esperado, o primeiro tempo foi marcado pelo equilíbrio*)
	At the second half, the match's overview hasn't changed. (*No segundo tempo, o panorama da partida não mudou*)
3	There were at least seven casualties, among which a Mozambican citizen, and 300 people were detained. (*Houve pelo menos sete mortos, entre os quais um cidadão moçambicano, e 300 pessoas foram detidas*)
	Over 300 people were detained for participating in vandalism actions. (*Mais de 300 pessoas foram detidas por participar de atos de vandalismo*)
4	The criminal organisation comprises various businessmen and a state congressman. (*A organização criminosa é formada por diversos empresários e por um deputado estadual*)
	According to the investigation, various businessmen and a state congressman formed the group. (*Segundo a investigação, diversos empresários e um deputado estadual integram o grupo*)
5	Other 8,869 won the tetrad and will earn R$ 356.43 each. (*Outros 8.869 fizeram a quadra e ganharão R$ 356,43 cada um*)
	At the tetrad 8.869 players won, the prize is R$ 356.43 each. (*Na quadra 8.869 apostadores acertaram, o prêmio é de R$ 356,43 para cada*)

Finally, although ASSIN2 relies on the same numerical scale as ASSIN to measure Semantic Similarity, it uses a different set of categories when it comes to Textual Inference. Also, some manual changes were made in ASSIN2 so as to have a more balanced data set for this task. This, however, does not affect our results, since we were aiming at the task of redundancy detection and, as such, depend on the Semantic Similarity scores only. Figure 3 shows the score distribution in both ASSIN and ASSIN2, illustrating the differences between both corpora.

3.2. Data pre-processing

As already mentioned, our source data sets assign, to every sentence pair, a similarity score in a scale ranging from 1 (no similarity) to 5 (high similarity). Since we are dealing with a binary classification problem (redundant $\times$ non-redundant), our first step was to map this scale to our target variable. To do so, sentence pairs classified as 4 or higher were assigned to the redundant class, with the remaining pairs being considered non-redundant. This division was based on a manual inspection of the data, whereby one of

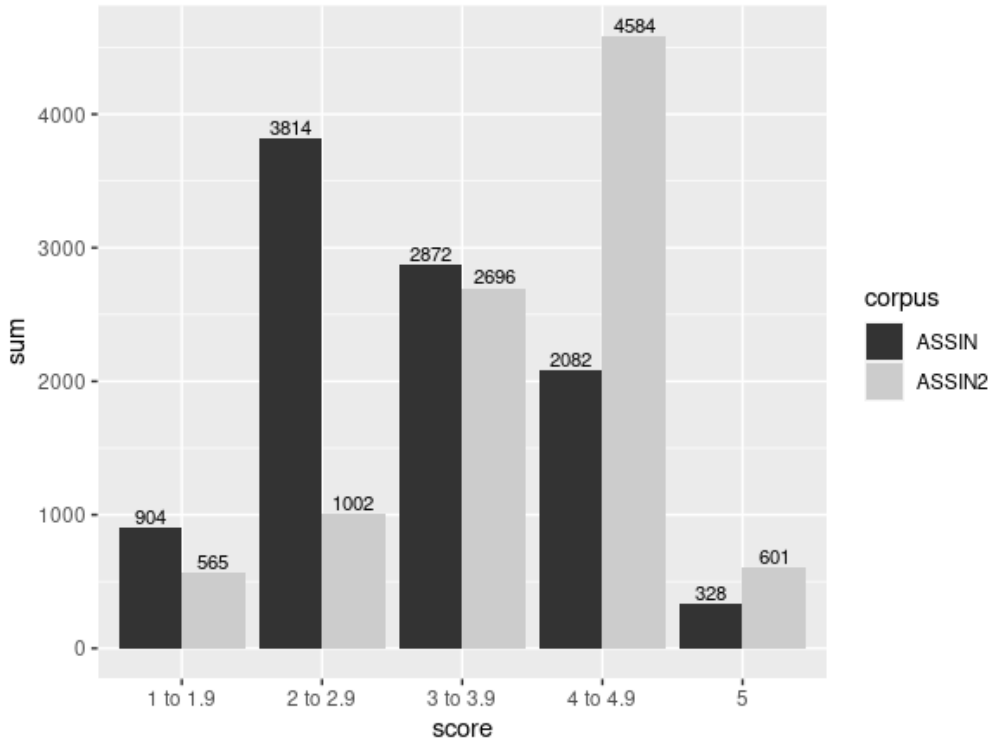

Figure 3. Score distribution across corpora.

the researchers went through a random sample of 50 sentence pairs with score 4 or higher, so as to determine whether the sentences in each pair could be considered similar.

Another important feature of these data sets is the fact that they come in three different files (see Section 3.1). These, however, and as already shown in Section 3.1, present different score distributions. To deal with this problem, training, testing and validation sets were grouped together, and a new split was randomly made (random_state seed = 25), where 20% of the data was left to testing, with the remaining 80% being used for training and validation purposes. Figure 4 shows the new data distribution across scores, for both training and testing sets, in ASSIN and ASSIN2.

3.3. Experimental setup

In this research, we used BERTimbau [Souza et al. 2020], a variation of BERT [Devlin et al. 2018] developed for Brazilian Portuguese, as our testing model. This choice was guided by results at SemEval 2019[5], a series of international natural language processing workshops aimed at advancing the current state of the art in semantic analysis, where BERT and its derivatives were widely used with interesting results. BERTimbau comes pre-trained in the BrWaC (Brazilian Web as Corpus) [Wagner Filho et al. 2018], and is available online for download[6].

Table 2 outlines the six different experiments that were carried out, corresponding to all possible combinations of two experimental variables: *data set* (ASSIN, ASSIN2 and the concatenated version of both) and *fine tuning* (true or false). The ultimate goal

[5]https://semeval.github.io/

[6]`https://github.com/neuralmind-ai/portuguese-bert`

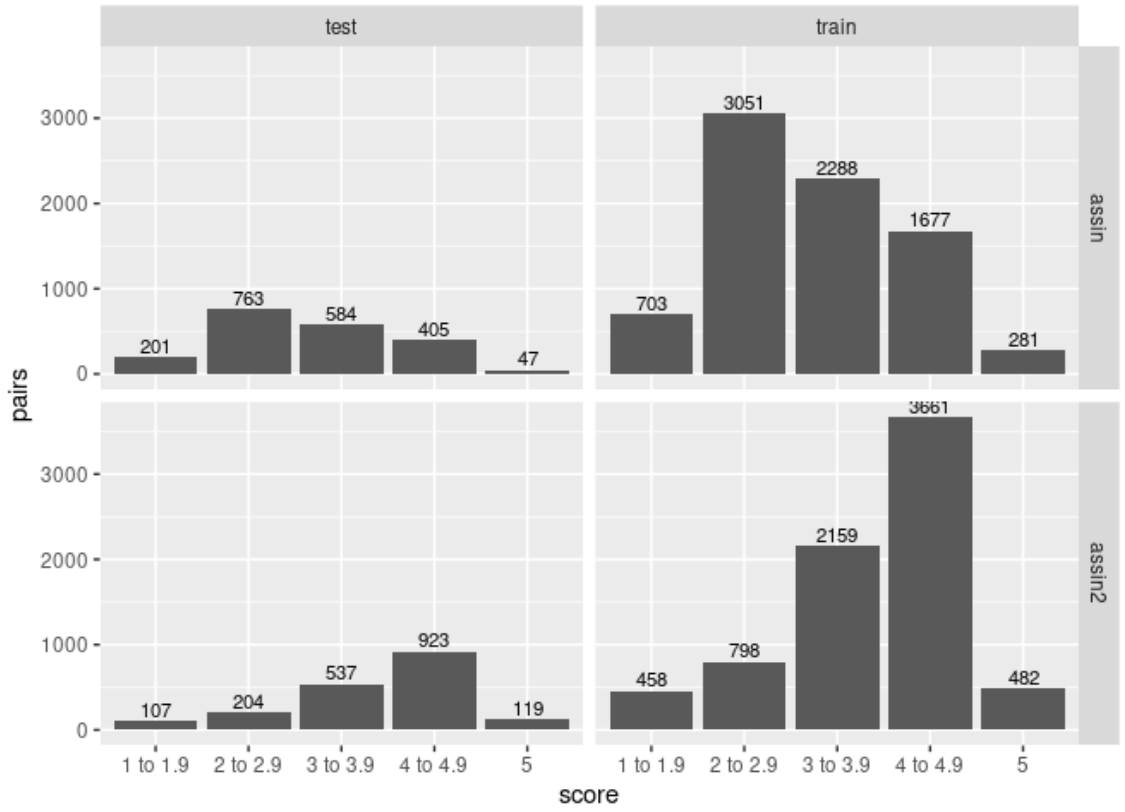

Figure 4. New score distribution in both corpora, in training and testing sets.

with this last variable was to verify if fine-tuning BERTimbau at the source corpus would increase its accuracy, specially when considering the extra time needed for this procedure. Table 3 summarises the configuration parameter values for the experiments. All experiments were performed using Google Colab with Python 3.7.13. For BERTimbau, we relied on the bert-base-portuguese-cased model, from the Hugging Face library[7]. Standard libraries, such as pandas, numpy and pytorch were also used, along with BERTimbau's tokenisation and classification modules.

Table 2. Executed experiments

Experiment	*Corpus*	*Fine Tuning*
1	ASSIN	TRUE
2	ASSIN2	TRUE
3	Concatenated version of ASSIN and ASSIN2	TRUE
4	ASSIN	FALSE
5	ASSIN2	FALSE
6	Concatenated version of ASSIN and ASSIN2	FALSE

BERTimbau's tokenisation module was run with *padding* = *True*, *truncation* = *True*, and maximum sentence length of 512 characters, returning Pytorch tensors. As its output, *input_ids* (tokenised instances), *attention_mask* and *token_type_ids* were selected.

[7]https://huggingface.co/neuralmind/bert-base-portuguese-cased

Finally, Pytorch's DataLoader, which is responsible for feeding the model with the training data, was configured to load batches of size 16, without shuffle. As suggested in Bert's article [Devlin et al. 2018], fine-tuned models were trained using three epochs, with 200 steps in the gradient direction per epoch. In all experiments, Cross Entropy was our loss function, along with the Adam optimizer.

Table 3. Configuration parameters for all experiments.

IDE = Google Colab
Python version = 3.7.13
Libraries = Hugging Face, Numpy, Pandas, Torch, sklearn and tqdm
BERTimbau model= bert-base-portuguese-case
Padding parameter = TRUE
Trucantion parameter = TRUE
Maximum sentence length = 512 characters
Load batches size = 16
Number of epochs = 3
Number of steps in the gradient direction per epoch = 200
Loss function = Cross Entropy
Optimizer = Adam
Number of folds for cross validation = 10

We then performed a 10-fold cross validation in the training set (*i.e.* the set corresponding to 80% of the original data). Since we were mainly interested in detecting redundancy in text, whereby two sentences are deemed as redundant or non-redundant, we adopted accuracy as our quality measure, so as to identify the overall amount of correctly classified examples, without accounting for which class produced the best results. The steps followed when performing the experiments were:

1. Importing standard libraries, sorting modules from Bert's library, tokenization module and pipeline module.
2. Setting control flags and hyperparameters.
3. Tokenizer definition from neuralmind/bert-base-portuguese-cased.
4. Data set splitting between training and testing sets.
5. Mapping the similarity scale to the target variable (redundant x non-redundant).
6. Tokenization of training and testing sets.
7. Running 10-fold cross-validation on the training set, measuring mean accuracy across folds, as well as each accuracy in each individual fold.

4. Results and Discussion

Table 4 shows the mean accuracy (in 10 folds) for our six experimental conditions. As it turns out, the best result was achieved by fine tuning BERTimbau at ASSIN, with this being the only condition where fine tuning the models resulted in better accuracy. Differences between the fine-tuning and no-fine-tuning groups ranged from -23% (ASSIN2) to 0.2% (with ASSIN). Interestingly, fine-tuning not only does not seem to raise accuracy but, as is the case with ASSIN2, even leads to a worse outcome. This is inline with some results regarding optimization difficulties found at the beginning

of training [Mosbach et al. 2020] or with what has been called "Catastrophic Forgetting" [Ede et al. 2022], whereby fine tuning a deep model might lead it to "forget" the main phenomenon learnt in its pre-training step.

Table 4. Mean accuracy in each corpus

Corpus	*No fine Tuning*	*Fine Tuning*
ASSIN	77.18%	77.40%
ASSIN2	67.92%	52.05%
ASSIN + ASSIN2	71.26%	57.88%

Accuracy varied to a great extent, from 52.05%, when fine-tuning the model in ASSIN2, up to 77.40%, with fine-tuned BERTimbau at ASSIN, an almost 49% increase over this baseline minimum value. In fact, the reduction in accuracy, from ASSIN to ASSIN2, was around 12% without fine tuning and almost 33% in the fine tuning group. This is a very puzzling result, given the fact that both data sets were supposed to have been built using a similar methodology and for the same tasks.

But then what aspects of these data sets could have led to such disparate results? The first point to be noted is that, although the data sets are meant to be applied to the same tasks and in the same language, there have been differences both in their source domain and construction. Hence, while ASSIN comes from journalistic texts, which may share a common structure, ASSIN2 was built from image captions, which are supposed to rely on a more simplified language (*i.e.* without presenting complex linguistic phenomena) and, in general, with shorter sentences.

Another point that might be relevant is the fact that, while ASSIN comprises sentence pairs originally produced in European and Brazilian Portuguese, ASSIN2 was created from a database that was automatically translated from English to Portuguese, and something may have been lost in translation, adding noise to the data and confusing annotators. In fact, evidence in favour of the existance of annotation problems came from a manual inspection of some sample sentence pairs classified as 4 or higher for semantic similarity. In this regard, whereas all pairs in ASSIN were confirmed redundant in this inspection, that was not the case with ASSIN2, as illustrated in Table 5, where one sees some clearly opposite sentences being assigned scores higher than 4, with the third pair being open for argumentation.

Finally, overall observed differences were found to be statistically significant[8]. Not surprisingly, a pairwise analysis showed no significant difference only between both versions of ASSIN (with and without fine tuning), and between ASSIN2 and the concatenation of both corpora without fine tuning, as revealed in a Tuckey post-hoc test, corrected for multiple comparisons.

5. Conclusion

Numerous articles have already demonstrated the benefits of distributional language models such as BERT and its derivatives (*e.g* [Hendrycks et al. 2020], [Wang et al. 2019], [Canete et al. 2020]). However, it is still possible to observe that there is room for improvement. In this research, one of BERT's derivatives – BERTimbau – was run in six

[8] $ANOVA(df = 5) = 88.39, p \ll 0.001$, at the 95% significance level.

Table 5. Some pairs scored as 4 or higher in ASSIN2 with opposite meanings

Score	*Pair*
4.085	A child is holding a water pistol. (*Uma criança está segurando uma pistola de água*)
	No child is holding a water pistol or being sprayed with water. (*Não tem nenhuma criança segurando uma pistola de água ou sendo pulverizada com água*)
4.05	A man and a woman are not shaking hands. (*Um homem e uma mulher não estão dando um aperto de mão*)
	A man and a woman are shaking hands. (*Um homem e uma mulher estão dando um aperto de mãos*)
4.2	A tiger is aimlessly roaming. (*Um tigre está andando sem rumo*)
	A tiger is roaming around the cage. (*Um tigre está andando em volta da gaiola*)

different experimental conditions, which correspond to all possible combinations of two experimental variables: data set (ASSIN, ASSIN2 and the concatenated version of both) and fine tuning (true or false).

Results showed that there are considerable differences in accuracy between the data sets. Moreover, fine tuning BERTimbau not only does not increase the model's accuracy, but sometimes make it worse. In the search for an explanation for these phenomena, we identified some aspects of the data sets that might have influenced these results. From this search, it became evident that the same model can present a considerable variation in accuracy across data sets, even when these are supposed to be very similar.

On this regard, we found that some annotation errors, along with the source of the data (which might bring some possibly different text structures to the scene) might be relevant to this question too. We, however, did not go any further so as to put these conjectures at test, leaving it for future research. Another venue for future investigation would be to verify whether this kind of result can be observed in corpora other than the ones we tested, along with languages other than Portuguese.

References

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ede, S., Baghdadlian, S., Weber, L., Samek, W., and Lapuschkin, S. (2022). Explain to not forget: Defending against catastrophic forgetting with xai. *arXiv preprint arXiv:2205.01929*.

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

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Marco, M., Luisa, B., Raffaella, B., Stefano, M., Roberto, Z., et al. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proc. SemEval*, pages 1–8.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Real, L., Rodrigues, A., e Silva, A. V., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C., Stanojević, M., et al. (2018). Sick-br: a portuguese corpus for inference. In *International Conference on Computational Processing of the Portuguese Language*, pages 303–312. Springer.

Rosu, R., Stoica, A. S., Popescu, P. S., and Mihăescu, M. C. (2021). Nlp based deep learning approach for plagiarism detection. In *RoCHI-International Conference on Human-Computer Interaction, Romania*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Tu, L., Lalwani, G., Gella, S., and He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. (2019). Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Yang, J., Zhou, K., Li, Y., and Liu, Z. (2021). Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.