





# BERT Self-Learning Approach with Limited Labels for Document Classification

Carlos Eduardo de Lima Joaquim<sup>1,2</sup>(✉)  and Thiago de Paulo Faleiros<sup>1</sup> 

<sup>1</sup> Departamento de Ciência da Computação, Universidade de Brasília,  
Campus Universitário Darcy Ribeiro, 70910-900 Brasília, Brazil  
`carlos.joaquim@live.com`, `thiagodepaulo@unb.br`

<sup>2</sup> Exército Brasileiro, Centro de Desenvolvimento de Sistemas, QGEx - Bloco G - 2o  
Piso, 70630-901 Brasília, Brazil

**Abstract.** The remarkable production speed of documents and, consequently, the volume of unstructured data stored in the Brazilian Government facilities requires processes that enable the capacity of classifying documents. This requirement is compliant with the existing archival legislation. In this sense, Natural Language Processing (NLP) stands as an important asset related to document classification, considering the reality of current document production, where there is a considerable number of unlabeled documentary samples. The Self-Learning approach applied to the BERT fine-tuning step delivers a model capable of classifying a partially labeled set of data according to the Requirements Model for Computerized Document Management Systems (e-ARQ Brazil). The developed model was capable of reaching a human-level performance, outperforming Active Learning and BERT in a series of defined confidence levels.

**Keywords:** Self-learning · BERT · Natural language processing

## 1 Introduction

With the advent of the information age, where information and communication technologies became an essential asset, the potential use of applications exploring the possibilities of obtaining useful and timely knowledge became real. Institutions can potentialize its documentary collection value, transforming it into a valuable asset.

There is an appreciable amount of information being produced in a daily basis, with a significant part of this collection becoming records related to legal and historical matter. The legal value regards the value a document has to produce evidence before the law, and the historical value concerns documents related to institutional origin, rights and objectives, its organization and development [1].

The authors of [2] declare that large volumes of information available and stored make it difficult to access for the right information at the right time. Stating that this situation might lead to the information explosions, in accordance with [23]. Along the same line of thought, [21 as cited in 2] affirms that, at the tactical level, poor information quality compromises decision making.

Inner statistics from the Brazilian Army show that it is possible to apply this method to more than 22 million documents, evaluating and classifying them according to the current Federal regulations. The needed classification of this documentary mass is the first step towards the direction of delivering efficiency while following the present regulations. That massive amount of data would take too long to be processed and assessed, if considering the possibility of scrutiny carried out exclusively by human hands.

This documentary production started to increasingly grow after Computerized Document Management System's (CDMS) initiative took place. The Federal classification model, named e-ARQ, was the chosen model to be applied to the documents, given that this archetype is the standard reference to CDMS in Brazil. Thus, this intended study is related to the pressing need to properly classify documents produced by the Brazilian Army, allowing correct treatment and full compliance with the requirements established by the Government.

In the present scenario, the shortage of labeled samples shall be considered as a premise. With this information being known, one major question when classifying documents, while following the supervised learning paradigm, is the existing need of a substantial number of labeled samples, in order to adequately generalize the model and make predictions of unseen samples, the supervised learning approaches does not become suitable as a deemed solution.

As the foregoing restriction regarding the number of labeled samples emerges as a limitation in several scenarios, the opposite turns out to be true. While labeled data is expensive to obtain, unlabeled data is essentially free in comparison [14]. It can be seen that creating large datasets to certain supervised learning problems requires a great deal of human effort, pain and/or risk or financial expense [11, 20]. This need for supervision poses a major challenge when we encounter critical scientific and societal problems where fine-grained labels are difficult to obtain [12].

In this line of thought, it can be seen in [20] that semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. The method can trace back to 1970s, and it attracts extensive attention since 1990s s [26 as cited in 13]

Considering that the size of modern real world datasets is ever-growing so that acquiring label information for them is extraordinarily difficult and costly [11, 18], deep semi-supervised learning is becoming more and more popular [13].

It becomes an attractive approach towards addressing the lack of data, once, in contrast with supervised learning algorithms, SSL algorithms can improve their performance by also using unlabeled examples. Additionally SSL algorithms generally provide a way of learning about the structure of the data from the unlabeled examples, alleviating the need for labels [20].

There are various fields within semi-supervised learning of which self-learning is one [17]. Throughout this study, it was expected to find language model that would dexterously classify the partially labeled dataset set according to the e-ARQ, reaching human-level performance [25] in the classification results in a set of documents belonging to the Brazilian Army.

It was envisaged to expand the use of BERT [5], and replace the fully supervised fine-tuning stage for a self-learning method, termed BERT-SL, expecting that the process surrounding BERT downstream tasks will successfully achieve suitable scores when evaluated.

Furthermore, from an organizational perspective, it was expected that the results coming from this research fulfill the objective of allowing adequate support to properly classify documents, assisting document evaluation teams to do their job, according to what is determined by [6].

The remaining of this article is organized as follows: Sect. 2 presents the works related to what was developed in this article. Section 3 presents the methodology used in the development of the work, describing how the procedures and experiments were performed aiming the evaluation. Section 4 describes the results obtained, according to the established parameters. Finally, in Sect. 5, the conclusion was carried out, as well as opportunities for future work were pointed out.

## 2 Related Work

This research's related work derives from different areas, here named: text classification, limited labeled data, semi-supervised learning, and self-learning. As specified by [19], the problem of learning accurate text classifiers from limited numbers of labeled examples, by using unlabeled documents to augment the available labeled documents, was addressed by showing that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents.

Their results showed EM to perform significantly better, mainly when there was little labeled data, though unlabeled data could throw off parameter estimation when one considered that the number of unlabeled documents was much greater than the number of labeled documents. The authors modulated the influence of the unlabeled data, in order to control the extent to which EM performs unsupervised clustering, by introducing a  $\lambda$  parameter into the likelihood equation, which decreased the contribution of unlabeled documents to parameter estimation.

Fragos, Belsis and Skourlas [7] focused on assessing the performance of two or more classifiers used in combination in the same classification task, classifying documents using two probabilistic approaches – Naïve Bayes and Maximum Entropy classification model – then combining the results of the two classifiers to improve the classification performance, using two merging operators, Max and Harmonic Mean.

The authors applied the  $\chi^2$  square test on the corpus and selected 2,000 higher ranked words for each category to be used in the maximum entropy model,

and evaluated the classification performance of the classifiers using micro-Recall ( $\mu Re$ ), micro-Precision ( $\mu Pr$ ) and micro-averaged F1 measure (*micro-F1*).

The Maximum Entropy model, presented by Fragos, Belsis and Skourlas [7], showed better performance than the Naïve Bayes classifier. Additionally, the two merging operators were used to combine results of the Naïve Bayes and SVM classifiers to improve performance, especially for the Recall rate. As results, it could be demonstrated that the merging operators do improve the performance, as indicated by the results for Micro-averaged F1 measure, that scored 0.90 and 0.91 for MaxC and HarmonicC operators respectively.

On the same pitch of [19], the use of self-learning and co-training is presented as a way to leverage the power of unlabeled data, together with labeled data, in [11]. The resulting work included the *TSentiment15*, an annotated Twitter dataset of 2015 comprising 228 million tweets without retweets and 275 million with retweets.

The authors evaluated the performance of self-learning and co-training and how it was affected by the amount of labeled data, the amount of unlabeled data and the confidence threshold, and, not only this, processed the available data as a batch and as a stream, showing that streaming achieved a comparable accuracy to the batch approach. The findings, in sentiment analysis, revealed that co-training performed better with limited labels, whereas self-training was best choice when significant amount of labeled data was available.

Iosifidis and Ntoutsis [11] used unigrams for self-learning, and unigram-bigrams and unigrams-SpecialF for co-training. In their experiments, although both self-learning and co-training benefited from more labeled data, when labeled data surpassed 40%, self-learning improved faster than co-training. When processing streaming, that revealed to be more efficient, history helped with the performance; notwithstanding the batch approach being better in terms of accuracy.

As avowed in [13], Li and Ye addressed issues related to generative model based schemes, that does not naturally work on discrete data. The authors bridged the idea of self-training and adversarial networks, to overcome their issues, by designing a Reinforcement Learning based Adversarial Networks for Semi-supervised Learning – RLANS – framework.

Howe, Khang and Chai [10] developed a comparative study on the performance of various machine learning (“ML”) approaches for classifying judgments into legal areas, using a novel dataset of 6,227 Singapore Supreme Court judgments and investigating how state-of-the-art NLP methods compared against traditional statistical models.

Dealing with small number of lengthy documents, the authors came to the conclusion that BERT models competed at disadvantage, because of the models’ inability to be fine-tuned on longer input texts, having LSA based linSVM outperforming both word-embedding and language models.

While Howe, Khang and Chai [10] have found limitations regarding the text length, citing it as a disadvantage, Sun *et al.* [24] conducted exhaustive experiments to investigate different fine-tuning methods of BERT on text classification

tasks and provided a general solution for BERT fine-tuning. This way they were able to reach new state-of-the-art results on eight widely-studied text classification datasets.

In [24], the problem related to the catastrophic forgetting was addressed as well. Considered a common problem in transfer learning, meaning that the pre-trained knowledge is erased while learning new knowledge [16 as cited in 24], it was managed by setting a lower learning rate, such as  $2e - 5$  on BERT, so it could overcome the catastrophic forgetting problem. Aggressive learning rates, as  $4e - 4$ , lead the training set to fail to converge.

The authors were able to bring off experimental findings, reporting that the top layer of BERT showed to be more useful for text classification, the appropriate decreasing learning rate allows BERT overcome the catastrophic forgetting problem, within-task and in-domain further pre-training can significantly boost its performance, and the most important finding considered to this work, BERT can improve the task with small-size data.

Meng *et al.* [18] used the label name of each class to train classification models on unlabeled data, waiving the use of any labeled documents. Towards achieving their goals, they took advantage of pre-trained neural language models for document classification. In this abstraction only the label name of each class was provided to train a classifier on purely unlabeled data.

In the same field, in [8] BERT model is compared with a traditional machine learning NLP approach that trains machine learning algorithms in features retrieved by the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm as a representative of traditional approaches. Experiments showed the superiority of BERT and its independence of features of the NLP problem such as the language of the text, adding empirical evidence to use BERT as a default technique in NLP problems [8].

### 3 Methodology

Considering that the main goal of this experiment is to assess the possible performance improvements that originate from the use of self-learning for downstream tasks, specifically text classification, the development of the methodology initially took place through the gathering of a specific dataset from a Military Organization (OM).

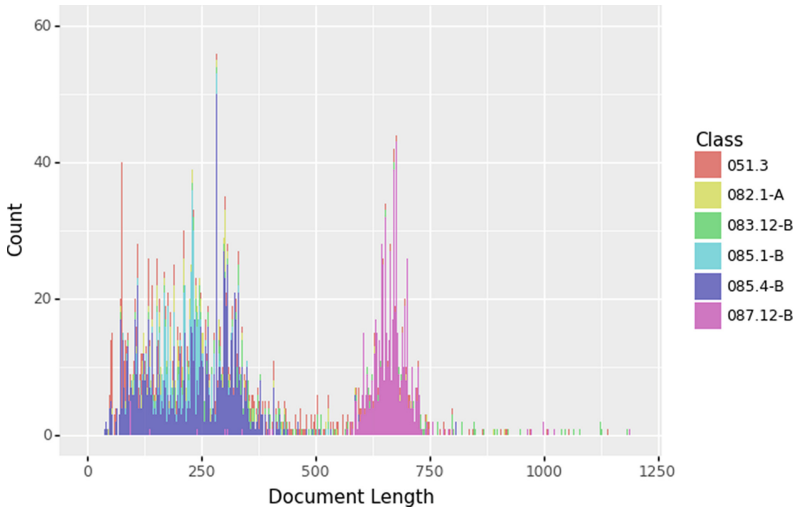
When it comes to the algorithm, initially a minimum set of labeled documents was trained in order to later classify the entire documents pertaining to the six chosen classes that encompass the following administrative actions, whose class id can be seen in Table 1: commendations, promotions, leaves, budget & finances, designations, and health. The process of labeling the unlabeled documents was based on the confidence level established as a threshold for the classification process.

Concerning the confidence level, the first experiments with classical BERT brought in f1-score 0.83 as one of the lowest results for specific classes, then a confidence level milestone was set having 0.82 as the basic confidence level for

unlabeled samples' classification. Subsequently, the threshold possibilities were expanded and additional confidence levels were considered as an option, starting from 0.6, and reaching 0.95.

This procedure continued iteratively until there were no more documents remaining to be classified at a specific confidence level. Afterwards the results were compared to the results that stem from another classification methods, Active Learning, associated with Logistic Regression model applied on a TF-IDF vectorized corpus, and BERT itself.

### 3.1 Data

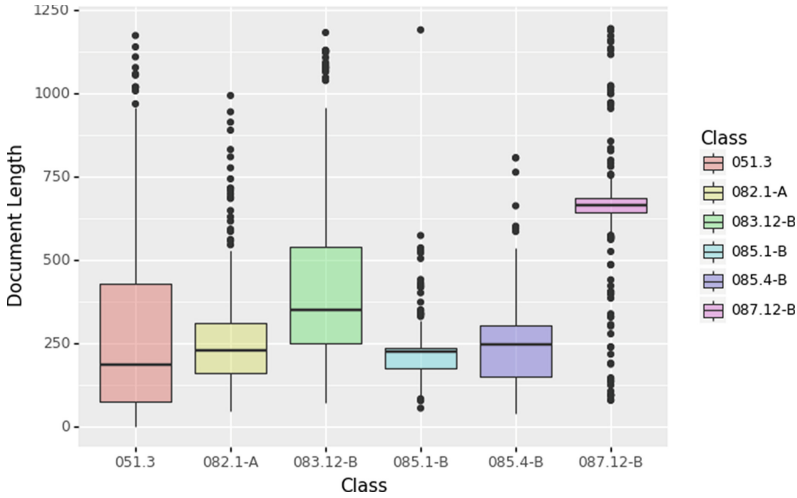


**Fig. 1.** Document length distribution

At the preprocessing stage, corrupted, drafts and non-processed documents were removed from the dataset, and all the corpus was converted to lowercase, along with manual preprocessing which comprehended: removing Portuguese stop words from the corpus using Natural Language Toolkit (NLTK) [3] library; Punctuation ablation; Numbers pertaining to itemization and object pronouns attached to Portuguese verbs removal; and conducting Lemmatization on the corpus using spaCy [9].

The corpus was then submitted to Ktrain [15] preprocessing methods. The resulting dataset had a  $5,940 \times 5,799$  dimensionality, with unbalanced class distribution, dispersed over six distinct classes, as seen in Table 1.

After data transformation, the resulting document length distribution, and class distribution can be seen in Fig. 1, and Fig. 2, allowing one to perceive the final length distribution of the dataset, and the final length distribution per class.



**Fig. 2.** Boxplot document length distribution per class.

**Table 1.** Class distribution

Class ID	Class description	Class size	%
085.4-B	Vacations or Medical Leave	2201	37.05
087.12-B	Verification of Medical Conditions	1352	22.76
051.3	Budget execution	926	15.59
085.1-B	Honorable Mention/Service Leave	810	13.64
083.12-B	Relocation	372	6.26
082.1-A	Promotion	279	4.70

In order to successfully establish a number of labeled documents, a classification tool, termed Document Classifier, was developed so the process of searching and classifying documents was conducted without further efforts regarding finding similar documents in the dataset.

### 3.2 BERTimbau

Throughout this work, finding an algorithm that could achieve state-of-the-art performance in Portuguese was a concern due to the existing linguistic bias in favor of languages that predominate on the areas where only major companies and research centers can afford training language models with billions of parameters on massive datasets [4].

The work of Souza *et al.* [22], BERTimbau, arose as the current answer to this need, delivering state-of-the-art results for downstream natural language processing tasks in Portuguese language, and will be simply referred to as BERT throughout this study. The pre-trained BERT [5] based model, BERTimbau,

was the model of choice, applied when running the fine-tuning step using the developed self-learning approach.

The model will be simply referred to as BERT throughout this study, having 12 layers, 768 hidden size, 12 attention heads, and 110M parameters. The maximum sentence length also observed BERT [5], following the  $S = 512$  tokens limit. The pre-trained BERT [5] based model, BERTimbau, was the model of choice, applied when running the fine-tuning step using the developed self-learning approach.

### 3.3 Self-Learning

Bearing in mind that this research can find in self-learning a solid answer to the problem of limited number of labeled documents in a dataset, it is intended to assess the resulting performance of the algorithm by selecting a specific percentage of samples from the total amount of labeled documents, starting with 3% and increasingly growing up this number to 30%.

**Table 2.** Four group sample distribution

Dataset size	Training size	Validation size	Unlabeled size	Test size
200	100	50	50	5740
594	269	146	179	5346
1188	540	291	357	4752
1782	810	437	535	4158

As a means to better achieve the objectives of this research, the data was split in four sets. In this fashion, during this work one will observe, as exhibited in Table 2, the four sets being used, notably the training set, validation set, unlabeled set, and test set.

As a way of exploring the possibilities of the self-learning process, experiments using the validation set as the unlabeled set were carried out as well. In this configuration, the datasets do not include the unlabeled set, and the validation set suffered prejudice by being classified according to the established threshold, having its samples moved from the validation set to the training set.

Further experiments included model where line 13 of the Algorithm 1 was suppressed. Yet, the sample distribution in the experiments was as follows: the methods were equally exposed to the same number of samples, thus reflecting a factual scenario, which commonly occurs in everyday life.

Observing the aforementioned data organization, in each experiment, the selected data was submitted to the self-learning Algorithm 1, until there were no more unlabeled samples in the corresponding set. After the initial training procedure, the full subset of documents were then submitted to the trained model, without any labels.



---

**Algorithm 1.** BERT Self-Learning Approach Pseudocode

---

**Input:**  $\mathcal{L}_{tr}$ : labeled training set;  $\mathcal{L}_v$ : labeled validation set;  $\mathcal{L}_{ts}$ : labeled test set;  $\mathcal{U}$ : unlabeled set;  $\delta$ : confidence threshold

**Result:**  $\mathcal{T}$ : labeled set;  $\Phi_f$ : final classifier

```

1:  $\mathcal{T} \leftarrow \mathcal{L}_{tr}, \mathcal{L}_v$ 
2: while ( $\mathcal{U}$  is not empty) do
3:    $\Phi \leftarrow$  new classifier
4:    $\Phi \leftarrow$  train new classifier on  $\mathcal{T}$ ;
5:   for  $i = 1$  to  $|\mathcal{U}|$  do
6:     if confidence of  $\Phi.classify(\mathcal{U}_i) \geq \delta$  then
7:        $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{U}_i$ , where  $\mathcal{U}_i$  is the  $i$ -th instance in  $\mathcal{U}$ 
8:       Mark  $\mathcal{U}_i$  as labeled;
9:     end if
10:  end for
11:  Update  $\mathcal{U}$  by removing labeled instances;
12: end while
13:  $\Phi_f \leftarrow$  train final classifier on  $\mathcal{T}$ 
14: return  $\mathcal{T}, \Phi_f$ ;

```

---

The resulting prediction for every sample, whose probabilistic classification was equal to or greater than the confidence level, allowed it to be incorporated, or not, in the labeled document selection of the training set for the next training session.

This process iteratively repeated its steps until all the documents in the  $\mathcal{U}$  dataset were considered labeled – to expound, the whole dataset received a classification equal to or greater than the defined confidence level.

**Downstream Task.** The first experiment had as main objective nothing more than the discovery of hyperparameters, *i.e.*, learning rate, number of epochs, and batch size, that would be expected to best perform when classifying the available dataset, bringing off satisfying results. It was possible to find the benchmark after having thoroughly tested all possible combinations of the values of the hyperparameters defined in Table 3.

**Table 3.** Benchmark model hyperparameters

Learning rate	Batch size	N <sup>o</sup> of Epochs
3e−5	2	4
4e−5	6	5
5e−5	8	6
6e−5	10	–
7e−5	–	–

Afterwards, series of self-learning experiments were conducted using the hyperparameters, and, then, more experiments were produced using classical BERT. The experiment was called BERT', being carried out having the data distributed, as presented in Table 4, and considered the most realistic scenario, once in real life there would be no more samples to feed classical BERT with.

**Table 4.** BERT' Data Distribution. In this distribution, as the unlabeled set was not used by the model, it was merged with the test set.

Dataset size	Training size	Validation size	Test size
200	100	50	5790
594	269	146	5525
1188	540	291	5109
1782	810	437	4693

## 4 Results and Discussion

In this section, the results of the experiments conducted using the self-learning approach during the fine-tuning step of BERT training are presented. The best hyperparameter experiment, where hyperparameter combinations would show the most prominent benchmark, registered that a  $4.00\text{E}-5$  learning rate, five epochs and batch size of two presented the best performance.

**Table 5.** Performance indicators (PI) of featured BERT-SL Classification Results. This table presents the best results for each dataset, outlined from the preceding experiments, against BERT' and Active Learning.

Dataset size	CL	F1-Score	PI	
			BERT'	AL
200	0.84	0.9771	0.0472	0.1330
594	0.82	0.9867	0.0165	0.0328
1188	0.90	0.9884	0.0146	0.0110
1782	0.90	0.9933	0.0121	0.0123

The ensuing experiments showed convincing results regarding the initial objective of this study, which was to find a self-learning method that could reach the human classification capacity.

As shown in Table 5, the use of self-learning also allowed the production of results comparable to BERT's, showing that it obtained excelling outcomes when

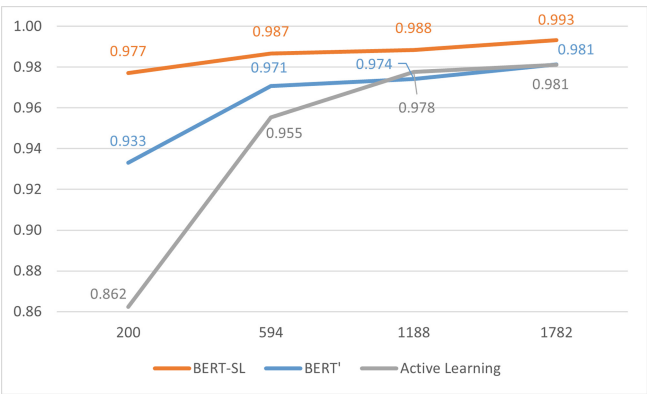
**Table 6.** Classical BERT classification results.

Dataset size	BERT' F1-score
200	0.9330
594	0.9708
1188	0.9743
1782	0.9814

compared to BERT, whose results are registered in Table 6, in several scenarios and confidence levels.

Despite of the highest score be related to the biggest dataset, the 1782-sample dataset, it is relevant to emphasize that the top gain came from the 200-sample dataset. The maximum gain coming from this experiment yielded a score 4.72% greater than BERT', and 13.30% beyond the active learning mark, as seen in Fig. 3, and detailed in Table 5.

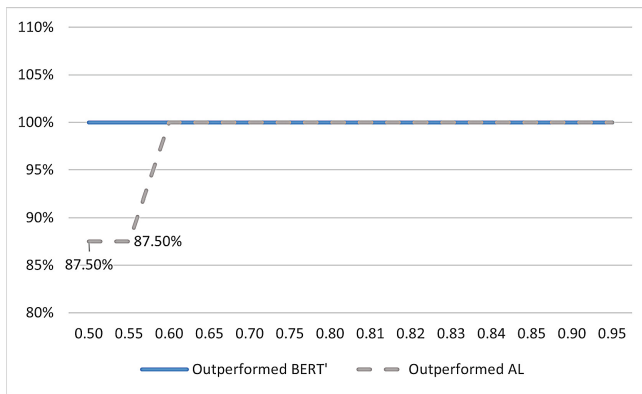
The same Fig. 3 allows one to perceive that BERT-SL, when working with the 594-sample dataset, provided a gain of 1.65% over BERT', and 3.28% over the active learning method. When dealing with 1188-sample dataset, the proposed method was able to deliver scores 1.46% better than BERT', and 1.10% greater than the active learning method. The gain got lower as the dataset size increased, reaching a gain of 1.21% over BERT' for the 1782-sample dataset, and 1.23% over the active learning method for the same dataset.



**Fig. 3.** BERT-SL overall performance, for every dataset, compared to BERT' and Active Learning experiments.

As presented in Fig. 4, throughout the undertook experiments, the language model resulting from the proposed method was able to outperform BERT' in every designed confidence level, surmounting the aforementioned method 100%

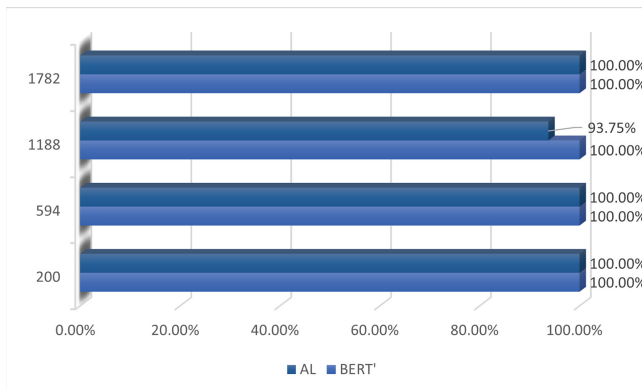
of the time, and surpassed the active learning process in every confidence level but 0.55 and 0.50, where BERT-SL was able to beat BERT' 87.50% of the time.



**Fig. 4.** Outperforming BERT experiments grouped by confidence level.

Considering the complete testing universe, Fig. 5 delineates the results. It remains clear that BERT-SL excelled BERT' in every experiment, and surpassed the active learning in almost all of them, being the only exception the tests where the 1188-sample dataset served as the input, when 93,75% of the experiments surpassed the active learning method.

It has been successfully demonstrated that, even when more samples were fed to classical BERT, BERT-SL showed to be able to achieve better scores, mainly when it comes to labeled samples shortage. The experiment, in Table 5, showed that BERT-SL was capable of achieving up to 4.72% better performance than the classical BERT, and 13.30% than the Active Learning.



**Fig. 5.** Percentage of BERT-SL experiments that outperform BERT'.

As described above, it was observed that, for experiments involving a set of two hundred samples, BERT-SL obtained a superior performance capable of, considering the aforementioned 22 million documents, estimating in 1,039,272 the number of documents adequately classified, when compared to BERT'; in 2,926,811 the number of documents adequately classified, when compared to the Active Learning method. The average gain for this set would be of 314,524 documents properly classified by BERT-SL, in comparison to the classical BERT.

## 5 Conclusion

Observing what was presented in the previous section, it is possible to infer, based on the results of this study, that the introduction of the self-learning approach in the fine-tuning stage allowed the improvement of BERT's performance, with emphasis on the target scenarios of the study that aims to treat the problem of scarcity of labels in the documental sets of the Brazilian Federal Government.

The self-learning approach, associated with BERTimbau, demonstrated to be a promising method regarding NLP classification tasks, showing better results than the classical BERT when the same number of samples is available to both methods, surpassing the traditional method in every single experiment following this setup.

The method showed outstanding results associated with datasets whose labels reach only 3% of the samples, showing an increasing performance, when compared to classical BERT, as the number of available labeled samples decreases. It was therefore possible to achieve suitable results while carrying out the experiment, making it possible to apply the method to other datasets.

## References

1. Nacional, A.: Gestão de documentos: curso de capacitação para os integrantes do sistema de gestão de documentos de arquivo siga, da administração pública federal. Course packet (01 2019), electronic Data (1 file: 993 kb)
2. Azemi, N., Zaidi, H., Hussin, N.: Information quality in organization for better decision-making. *Int. J. Acad. Res. Bus. Soc. Sci.* **7** (2018). <https://doi.org/10.6007/IJARBSS/v7-i12/3624>
3. Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python* (2009). <https://nltk.org/book>
4. Castro, N.F.F.d.S., da Silva Soares, A.: Multilingual transformer ensembles for portuguese natural language tasks (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
6. Exército Brasileiro: Instruções gerais para avaliação de documentos do exército (10 2019), eB10-IG-01.012
7. Fragos, K., Belsis, P., Skourlas, C.: Combining probabilistic classifiers for text classification. *Procedia-Soc. Beh. Sci.* **147**, 307–312 (2014)
8. González-Carvajal, S., Garrido-Merchán, E.C.: Comparing Bert against traditional machine learning text classification (2021)

9. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: industrial-strength natural language processing in Python (2020). <https://doi.org/10.5281/zenodo.1212303>
10. Howe, J.S.T., Khang, L.H., Chai, I.E.: Legal area classification: a comparative study of text classifiers on Singapore supreme court judgments (2019)
11. Iosifidis, V., Ntoutsi, E.: Large scale sentiment learning with limited labels. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 2017, New York, NY, USA, pp. 1823–1832. Association for Computing Machinery (2017). <https://doi.org/10.1145/3097983.3098159>, <https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3097983.3098159>
12. Jean, N., Xie, S.M., Ermon, S.: Semi-supervised deep kernel learning: regression with unlabeled data by minimizing predictive variance (2019)
13. Li, Y., Ye, J.: Learning adversarial networks for semi-supervised text classification via policy gradient. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1715–1723 (2018)
14. Liang, P.: Semi-supervised learning for natural language. Ph.D. thesis, Massachusetts Institute of Technology (2005)
15. Maiya, A.S.: ktrain: a low-code library for augmented machine learning. CoRR abs/2004.10703 (2020), <https://arxiv.org/abs/2004.10703>
16. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. *Psychol. Learn. Mot.* **24**, 109–165 (1989)
17. McEntee, E.: Enhancing partially labelled data: self learning and word vectors in natural language processing (2019)
18. Meng, Y., et al.: Text classification using label names only: a language model self-training approach (2020)
19. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **39**(2), 103–134 (2000)
20. Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J.: Realistic evaluation of deep semi-supervised learning algorithms (2019)
21. Redman, T.C.: Improve data quality for competitive advantage. *MIT Sloan Manage. Rev.* **36**(2), 99 (1995)
22. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20–23 (2020, to appear)
23. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. *Commun. ACM* **40**(5), 103–110 (1997)
24. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune Bert for text classification? (2020)
25. Wolf, F., Poggio, T., Sinha, P.: Human document classification using bags of words, August 2006
26. Zhu, X.J.: Semi-supervised learning literature survey (2005). last modified on 19 July 2008